CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Mining Drug Resistance Patterns In Mycobacterium Tuberculosis Using Bioinformatics Approaches

by

Anam Tariq

A dissertation submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Health and Life Sciences
Department of Bioinformatics and Biosciences

2024

Mining Drug Resistance Patterns in Mycobacterium Tuberculosis Using Bioinformatics Approaches

By

Anam Tariq

(DBI163002)

Dr. Amir Hussain, Professor

Edinburgh Napier University, UK

(Foreign Evaluator 1)

Dr. Muhammad Sajid Hussain, Professor

Cologne Center for Genomics, University of Cologne, Germany

(Foreign Evaluator 2)

Dr. Sahar Fazal

(Research Supervisor)

Dr. Syeda Marriam Bakhtiar

(Head, Department of Bioinformatics and Biosciences)

Dr. Sahar Fazal

(Dean, Faculty of Health and Life Sciences)

DEPARTMENT OF BIOINFORMATICS AND BIOSCIENCES

CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

ISLAMABAD

2024

With love, First and foremost I would like to dedicate this work to the Almighty ALLAH, Who blessed me with the courage and commitment to complete this work. I extend my deepest thanks to my parents, who have consistently prayed for me, showered me with love, and provided unwavering support throughout my life. My sister who has been my unwavering companion in every stage of life. My husband's constant inspiration has been a driving force behind my efforts, and I am truly grateful. To my beloved daughter, Zimal Kashif. I am thankful for their presence and encouragement.

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD

Expressway, Kahuta Road, Zone-V, Islamabad
Phone:+92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: https://www.cust.edu.pk

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the dissertation, entitled "**Mining Drug Resistance Patterns in Mycobacterium Tuberculosis using Bioinformatics Approaches**" was conducted under the supervision of **Dr. Sahar Fazal**. No part of this dissertation has been submitted anywhere else for any other degree. This dissertation is submitted to the **Department of Bioinformatics & Biosciences, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Biosciences**. The open defence of the dissertation was conducted on **July 10, 2024.**

Student Name : Anam Tariq (DBI163002)

The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

(a) External Examiner 1: Dr. Fariha Hassan
Professor
QAU, Islamabad

(b) External Examiner 2: Dr. Amjad Ali,
Associate Professor
ASAB, NUST, Islamabad

(c) Internal Examiner : Dr. Erum Dilshad
Associate Professor
CUST, Islamabad

Supervisor Name : Dr. Sahar Fazal
Professor
CUST, Islamabad

Name of HoD : Syeda Marriam Bakhtiar
Associate Professor
CUST, Islamabad

Name of Dean : Dr. Sahar Fazal
Professor
CUST, Islamabad

AUTHOR'S DECLARATION

I, **Anam Tariq** (Registration No. **DBI163002)**, hereby state that my dissertation titled, '**Mining Drug Resistance Patterns in Mycobacterium Tuberculosis using Bioinformatics Approaches**' is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

(**Anam Tariq**)

Dated: 10 July, 2024

Registration No : DBI163002

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the dissertation titled "**Mining Drug Resistance Patterns in Mycobacterium Tuberculosis using Bioinformatics Approaches**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete dissertation has been written by me.

I understand the zero-tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled dissertation declare that no portion of my dissertation has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled dissertation even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized dissertation.

(**Anam Tariq**)

Dated: 10, July, 2024

Registration No : DBI163002

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this dissertation:-

1. **Tariq, A.**, & Fazal, S. (2024). "Exploring Genomic Patterns to Identify Drug-Resistant TB: A Comprehensive Study of Age, Gender, Lineage, and Outcome." Advancements in Life Sciences, Vol No. 11, Issue No. 2, PP. 354-361.

(Anam Tariq)
Registration No: DBI163002

Acknowledgement

All gratitude is directed towards the Almighty Allah, who is the epitome of generosity and empathy, and His Holy Prophet Mohammad (Peace be upon Him), the most exemplary and illustrious individual ever to grace the Earth. He continues to be a guiding light and a source of wisdom for all of humanity. It is through his blessings that I have found the strength and skill to accomplish this objective. Secondly, it fills me with a sense of pride to share some thoughts about my esteemed research mentor, Dr. Sahar Fazal, who consistently and effectively instilled a spirit of curiosity when it comes to research. Her enthusiasm, assurance, intellect, and unwavering support throughout every phase of this entire project have empowered me to achieve my objectives. I must emphasize that without her guidance and generous dedication, this task would have been insurmountable. I am profoundly thankful to my family, and no words can truly convey the depth of my gratitude. My parents, whose unwavering dedication throughout my life has paved the way for my success, deserve special recognition. In particular, my father's constant motivation has driven me to achieve my goals, and he continues to enrich me with the invaluable gift of education. My sister has been a consistent source of support in every aspect of life, and her presence has been invaluable. Last but certainly not least, my husband has been a pillar of support, making this journey much more manageable. I owe a heartfelt appreciation to all of them for their unparalleled support, love, and prayers.

(Anam Tariq)

Abstract

Drug-resistant tuberculosis poses the most significant challenge to the global eradication efforts against tuberculosis. Healthcare institutions possess extensive repositories of data, which can prove highly valuable in forecasting the implications of drug responses. TB Portals is a huge data consortium comprising clinical demographic, bacterial genomic data and drug resistance data from TB patient cases from 15 countries throughout Europe, Asia, and Africa. The acquired data was of three types i.e demographic data, drug resistance data and genomic data. This study also utilized explanatory data analysis to investigate the impact of demographics, treatment outcomes, and genomic mutations on drug-resistant TB using patterns in data. The dataset consisting of 2,602 entries was filtered using pattern recognition techniques to identify significant features. The drug resistance data was clinical/phenotypic results of invitro drug susceptibility testing. The clinical isolated were tested against 18 anti TB first line and second line drugs. The aim of this study is to apply association rule mining on drug susceptibility data and identify the interesting pattern and further validate the patterns through genomic data. After preprocessing the association rule mining was applied 1970 entries. A whole genome sequence pipeline (ARIBA) was applied on genomic data to identify the variants. These variants were relevant to the patterns derived from clinical and demographic data. Around 1041 genomes sequences of drug resistant TB isolates were acquired from NCBI. Furthermore, the identified genes were subjected to functional enrichment analysis and hub gene identification techniques. Drug-resistant TB strains, including multidrug-resistant (MDR) and extensively drug-resistant (XDR) types. The results revealed that XDR and MDR non-XDR TB were prevalent types of drug resistance. Males exhibited a higher susceptibility to both XDR and MDR non-XDR TB. The association rules representing MDR were ranked highest which means Rifamipicin and isoniazid resistance was most abundant.Among XDR isolates fluoroquinolones, capreomycin (CAP) and streptomycin (STR) and ethambutol(EMB) resistance was most prevalent in addition to isoniazid and rifampicin's .The study aimed to identify novel gene variants in XDR and MDR strains associated with drug resistance. ARIBA identified 47

genes involved in antimicrobial resistant pathways ,6 were exclusively MDR unique novel variants, 28 among the observed variants in 31 commonly occurring genes. for XDR and MDR. Around 12 variants were associated with XDR/MDR strains, lacks literature evidence for their involvement in antimicrobial pathways specific to M. tuberculosis . Enrichment analysis revealed their association with antimicrobial resistance, RNA binding, ribosomal proteins, cell wall biogenesis/degradation, and enzymatic activities. Specific novel mutations in the *rpoC* gene, such as G332R, F452C, V864I, Q887K, and A898T were identified which were associated with enhanced fitness and drug resistance, suggesting further research on genes like *rpoB, rpoC, rpoZ*, and *rpoA* with demonstrated relevance to *M.tuberculosis* drug resistance and novel variants.These findings introduce a novel set of therapeutic targets specific to MDR and XDR TB types.

Contents

List of Figures

List of Tables

Abbreviations

AECs	Airway Epithelial Cells
AMK	Amikacin
AMR	Antimicrobial Resistance
AGs	aminoglycosides_injectible_agents
ARIBA	Antimicrobial Resistance Identification by Assembly
ARM	Association Rule Mining
BCG	Bacillus Calmette Guerin
CAP	Capreomycin
CARD	Comprehensive Antimicrobial Resistance Database
CS	Cycloserine
CSV	comma-separated values
DM	Data Mining
DST	Drug Susceptibility Testing
EDA	Explanatory Data Analysis
EMB	Ethambutol
ETH	Ethionamide
FDA	Food and Drug Administration
FLD	First Line Drugs
FQs	Fluoroquinolones
GO	Gene Ontology
GWAS	Extensive Genome-Wide Associate Studies
INH	Isoniazid
KAN	Kanamycin

KEGG	Kyoto Encyclopedia of Genes and Genomes
LVX	Levofloxacin
LZD	Linezolid
MDR	Multi Drug Resistance
ML	Machine Learning
MOX	Moxifloxacin
NCBI	National Center for Biotechnology Information
NIAID	National Institute of Allergy and Infectious Diseases
NKs	Natural Killer Cells
OFL	Ofloxacin
PAS	p_aminosalicylic_acid
PDB	Protein Data Bank
PTH	prothionamide
PZA	Pyrazinamide
RIF	Rifampicin
SAM	Sequence Alignment/Map
SLD	Second Line Drugs
STR	Streptomycin
TB	Tuberculosis
WHO	World Health Organization
XDR	Extensive Drug Resistance

Chapter 1

Introduction

1.1 Introduction

Tuberculosis (TB) is an ancient disease that has affected mankind for more than 4,000 years. TB is characterized as the one of the most virulent disease with highest death toll among other infections worldwide [1]. Mostly TB would affect people in their reproductive age, but other age groups are at risk too. The incidence of infection is followed by vigorous progression of the disease. It mainly affects lungs but other organs such as kidney, spine, and brain may get infected. In most of cases the bacteria remain dormant, but among one tenth of the cases it causes infection which is called active TB. The causative agent of this disease, Mycobacterium (*M. tuberculosis*), is capable of retaining its activity in tissues of a strong person, which results in a delayed diagnosis until it is transmitted to other hosts. It is pertinent to mention that late diagnosis also delays the treatment of the disease. The disease spreads from a person to another via air through inhalation of contaminated droplets. The pathogen propels into the air when a lung TB career cough, sneeze or spit and the other person who inhale these germs, will become infected [2].

M. tuberculosis is a pathogenic bacteria that belongs to the family of Mycobacteriaceae distinguished by lipid rich cell walls which help protect it from host

immune responses and survive stressful environments [3]. Mainly, *M. tuberculosis* attacks the lungs where it can be retained throughout the lifetime as an inactive latent form; however, upon reactivation it drives active transmission of the infection causing active TB [4]. *M. tuberculosis* first disarms the innate immune response as first line of defense by disrupting the activity of its major players including neutrophils, natural killer (NK) cells, macrophages, mast cells, dendritic cells, and airway epithelial cells (AECs) [5]. Genome size of *M. tuberculosis* is around 4.4 million base pairs containing over 4,000 genes, which is relatively large in comparison with other bacterial species [6]. However, the progression of TB is majorly associated with PE/PPE family genes, enoyl-CoA hydratases, and mce3 genes of *M. tuberculosis*, which play pivotal roles in host-pathogen interactions [7].

According to the WHO's Global TB Report 2022, 10.6 million people were diagnosed with TB in 2021 with a rise in the incidence rate of TB as compared to the past years [8]. TB is most common in developing countries reporting over 80% of cases and deaths. Recent reports estimated the largest number of new TB cases in the South-East Asian region leads with 46% of cases followed by African region and Western Pacific regions with 23%, 18% reported cases respectively. Thirty countries are considered as high TB burdened countries, as they have a total of 87% of cases reported. Among those thirty, eight countries are leading with two-third of the TB cases including Pakistan at fifth position with 5.8% of cases . Around the world, in total 10.6 million people suffered from TB disease in 2021 with 6 million men, 3.4 million women and 1.2 million children [1].

Prevention from fatal infectious diseases like TB depends on the antibiotics that are the antimicrobial drugs used to fight against pathogens causing deadly diseases. Antibiotics have played a critical role in revolutionizing the healthcare industry by treating a number of pathogenic infections and reducing the overall mortality rate [9]. However, inappropriate use of antibiotics carry widespread adverse effects, inducing antibiotic resistance by altering the composition of the pathogen which gives rise to new resistant strains against commonly used antibiotics [10]. Moreover, bacterial resistance occurs when bacterial cells acquire resistance against bactericidal effects for infections caused by bacteria including *M. tuberculosis* [11].

The effective and permanent treatment of TB can last from 6 to 9 months during which various antibiotics are prescribed to the patient. Ten drugs are globally approved for TB treatment by the U.S. Food and Drug Administration (FDA).The antibiotics which are considered as first-line drugs in treatment regimens are isoniazid (INH), rifampin (RIF), pyrazinamide (PZA), and ethambutol (EMB) are taken for about 2 months. In this phase of treatment, the rapidly growing bacteria are targeted and a successful first line treatment leads to eradication of clinical symptoms. But some of the bacteria develop resistance against drugs resulting in relapse and the spread of the disease. The advent of multidrug resistance TB (MDR-TB), i.e. which is resistant to at least isoniazid (INH) and rifampicin (RIF), is major concern, because it involves the usage of second-line drugs that are comparatively problematic to procure and are much more toxic and expensive than first line drugs (FLDs). The second, yet essential stage of TB treatment, is the 4 months continuation phase, in order to kill the stubborn or slow growing strains *M. tuberculosis*, which has produced resistance against FLDs. Common second line drugs (SLDs) are Fluoroquinolones and some injectable anti TB drugs [12].

There are numerous strains of *M. tuberculosis* that are drug resistant, majorly F15/LAM4/KZN and Beijing strains reported to be most common in MDR (multidrug resistant) and XDR (extensively drug resistant) outbreaks [13]. A recent study confirmed that MDR and XDR being the most frequent types of drug resistance in TB, are linked with poor treatment success rates which is around 30% for XDR and 54% for MDR [14]. Mainly, there are seven distinct lineages of *M. tuberculosis* worldwide. However, F15/LAM4/KZN and Beijing sublineages are the most dominant and significantly associated with drug resistance, majorly targeting younger ages [13, 15]. Bacterial antibiotic resistance in *M. tuberculosis* strains is predominantly linked to chromosomal mutations of the selected genes typically harboring various mechanisms including target disruption via enzymatic modification, changes in efflux pumps, and overexpression of the target [16]. RIF resistance is a public health dilemma that occurs due to the mutated rpoB gene. Most frequent mutations inducing RIF resistance occur at codons 435, 445, and 450 [17]. Similarly, anti-TB drug isoniazid (INH) inhibits mycoloic acid present in

bacterial cell walls. S315T mutation occurring at codon 315 of katG gene is the most common mutation causing INH resistance in patients with TB [18]. When TB becomes MDR, which means resistant to the two most potent drugs i.e. rifampin and isoniazid, the risk of treatment failure also increases [19]. Second line injectable drugs known as fluoroquinolones are used to treat such cases by preventing the synthesis of bacterial DNA. However, resistance of TB isolate against rifampin, isoniazid, and fluoroquinolones corresponds to XDR which is the most dangerous version of MDR linked with an increased mortality rate [20]. Moreover, mutated pncA gene has been reported to be crucial in determining the clinical outcomes when studied with other factors such as age and the utilized treatments procedures [21].

To get a deeper insight into the underlying genomic patterns associated with drug-resistant TB, various demographic features majorly lineage, outcome, age, gender, and type of resistance serve as key determinants [22]. Analysis of gender distribution among TB patients showed that the majority of reported cases were males with 52-53% higher risk in comparison with females, which increases with age [23].

TB remains a major health problem worldwide. It has been reported that MDR being the most common type of resistance is frequently found in the new cases of TB yet the available therapies only provide $55-65\%$ of survival rate despite completing a tedious and time-consuming treatment [24]. Moreover, the currently existing TB treatments have certain limitations including treatment complexity, time duration, toxicity, and the increasing resistance to the anti-TB drugs. Additionally, it is difficult to eliminate TB, as *M. tuberculosis* has the ability to persist within the host for a lifetime without causing the infection [25]. The available therapies for such cases including isoniazid monotherapy carry detrimental effects for instance, isoniazid may induce polyneuropathy, hepatitis, and jaundice [26]. Although, several solutions have recently emerged to target TB with lesser complexities; unfortunately, clinical research related to the therapies is still under process and might take a long time in declaring those treatments as clinically useful regimens [27]. To eradicate the epidemic of TB and its drug resistance, an intensified research with novelty and innovation and extensive genome-wide

association studies (GWAS) are required to be done to understand and develop more effective interventions to detect, cure and prevent it [28]. Since the new anti-TB drug candidates that are already in preclinical development phase are lesser in number and due to the emerging resistant strains, there is a dire need to introduce new drugs by considering the genomic patterns responsible for drug resistance [29]. This can be facilitated through exploratory data analysis (EDA) which uses pattern recognition as a fundamental activity to search for the patterns and trends in the isolated samples, in order to shortlist them based on their clinical features followed by visual inspection of the selected features [30]. Moreover, it will provide an insight for determining the antimicrobial resistance (AMR) genes that are specific to each subtype of the isolated samples or strains of *M. tuberculosis*, in turn helping to investigate novel anti-tubercular therapeutic biomarkers that can be targeted to each specific subtype of resistant *M. tuberculosis* strain in TB patients [31].

The field of medical science is considered full of information but still it lacks in terms of knowledge. A huge amount of healthcare data is present online or in systems of medical facilities. Nevertheless, there is a shortage of active tools to determine concealed yet useful relations and patterns in data. Data Mining (DM) have many applications in business and science. Significant knowledge is derived by applying of data mining techniques in clinical datasets [32].

Data Mining is the field of computer science in which we discover information from huge datasets and derive patterns and models. Datamining also referred Knowledge Discovery in Database (KDD), as involves techniques from other areas as machine learning, statistics, artificial intelligence, database sets, pattern recognition and visualization. There are different tasks which are performed in datamining enlisted as: classification, estimation, prediction, association rule mining, clustering, description, and visualization [33]. The latest in vivo techniques for predicting, patterns and deriving propositions in bioinformatics have been evolved. The data mining process has a lot of applications in bioinformatics comprising of gene finding, protein function domain detection, function motif detection and protein function inference. The bioinformatics data banks such as the Protein

Data Bank (PDB) have millions of records besides this a large amount of clinical data is generated on daily basis and is available on public and private platforms. The data mining methods like clustering, classification, association rules mining (ARM) have been successfully applied on public health data to discover interesting patterns [34]

The purpose of this study is to apply pattern identification and data mining techniques to a clinical dataset of *M. tuberculosis* in order to analyze the impact of gender, age of onset, lineage, type of resistance, and treatment outcomes on the identification of specific types of drug resistance associated with treatment failure. Additionally, the study employs the whole genome sequencing (WGS) pipeline to identify single nucleotide polymorphisms (SNPs) that result in new mutations in drug-resistant *M. tuberculosis* strains, specifically those leading to multidrug-resistant (MDR) and extensively drug-resistant (XDR) types. This analysis aims to discover novel therapeutic targets that can effectively combat these resistant and pathogenic strains of *M. tuberculosis*, particularly MDR and XDR types A major challenge in TB treatment is prediction of the clinical response to antibiotics for each individual due to bacterial resistance against specifically first line antimicrobial drugs. Data mining techniques can be used to identify patterns from clinical databases. These patterns will be useful in reviewing the mutations responsible for the resistance.

1.2 Research Problem

The ongoing challenge of drug-resistant tuberculosis (TB) poses a considerable hurdle in global TB eradication efforts. Despite the presence of extensive data archives within healthcare institutions encompassing clinical, genomic, and drug resistance data, there is a notable absence of reliable tools for precisely forecasting the consequences of drug treatments, including drug resistance patterns. This knowledge gap constrains our capacity to create customized therapies and effectively address the proliferation of drug-resistant TB variations.

1.3 Research Objectives

This study was intended to:

1. To understand demographic characteristics of TB patients and their association with treatment outcomes, lineage, and drug resistance using Exploratory Data Analysis (EDA).

2. To uncover significant MDR and XDR patterns in antimicrobial susceptibility testing data with association rule mining.

3. To analyze genomic sequences of drug-resistant TB isolates and identify existing and unique mutations to establish relationship with pattern through datamining.

4. To perform Functional Enrichment Analysis to understand gene functions and assess their functional significance in TB drug resistance.

5. To identify hub genes and construct a network to explore molecular interactions related to TB drug resistance.

1.4 Research Philosophy

A number of potentially curative antibiotics are present but still, TB continues to cause sickness and mortality at alarming rates globally , particularly in developing countries. The instances of MDR-TB or XDR-TB arises in one of two ways: (1) firstly, when a person gets infected with MDR or XDR TB strain. Secondly resistance may develop in TB patients from misuse or mismanagement of anti-TB drugs. MDR-TB and XDR-TB typically demands a significantly prolonged treatment period (up to two years), in contrast to the standard regimen for drug-susceptible TB. Normal drug sensitive TB infections can be cured with the first-line anti-TB drug regimen. However, managing DR-TB is challenging and less

promising, resulting in the ongoing persistence of the TB pandemic. The second-line anti-TB drugs, recommended for MDR- and XDR-TB, are toxic, costly and are relatively less effective as compared to the first-line drugs. Moreover, the battle against TB is complicated due to factors like HIV co-infection, the influence of COVID-19, patient compliance issues, and suboptimal treatment approaches across different regions of the globe.

Since the completion of sequencing of *M. tuberculosis* genome in 1998, which is complex as it contains approximately 4000 genes making is challenging to understand molecular biology of the bacterium. Approaches utilizing whole genome sequencing offer more advanced insights into mutations-based genotyping, profiling drug resistance, and detecting patterns of transmission. Comparative genomics analysis have effectively identified numerous new mutations in *M. tuberculosis* that are associated with resistance or adaptations. Although it is believed that all mutations leading to drug resistance would result in a competitive fitness cost relative to susceptible strains. research has revealed that clinical strains frequently harbour mutations with low or no fitness cost and that the fitness cost of other mutations can be offset by compensatory mutations. The comprehensive understanding of the *M. tuberculosis* genome has helped researchers to identify a subset of genes that are crucial for both in vitro and in vivo contexts. Since M. TB genome is sequenced, numerous small molecules possessing strong efficacy against both drug-susceptible (DS) and drug-resistant (DR) M. TB strains have been revealed, along with the elucidation of their particular targets. Certainly, many researchers have shifted their focus towards newly identified drug targets in *M. tuberculosis*, moving away from the traditional targets of current TB antibiotics to overcome drug resistance. However, a few of the drugs that target these newly identified targets have shown problems like toxicity, limited effectiveness in vivo, or short elimination half-life. Despite significant efforts made to introduce more effective anti-TB drugs, only three new medications with novel mechanisms have been approved since 2013, over a span of more than fifty years, and these are associated with serious side effects. Therefore, introducing new, potent anti-TB compounds

into the drug development pipeline could potentially accelerate the discovery of groundbreaking TB treatments.

The sequence analyses of *M. tuberculosis* strains are also providing new insights into the ongoing evolution of *M. tuberculosis* during infection, treatment and the acquisition of drug resistance. These results can also be used to develop more effective strategies for deploying existing drugs, such as by analysing drug resistance mutations in patient-derived populations of *M. tuberculosis* to predict the drugs that might be most clinically effective for a particular patient. In conclusion, systems mapping uncovers complex regulatory systems that have evolved to aid the organism in surviving within the host. Disturbing these systems could potentially open novel paths for drug discovery.

1.5 Research Hypothesis

As the worldwide health crisis posed by multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB strains persists. The objective of the study is to reveal previously undiscovered genetic variants that contribute to antimicrobial resistance.By utilizing of data mining techniques and conducting a genomic analysis of clinical datasets containing MDR and XDR TB strains, it is aimed to elucidate the genetic underpinnings of resistance mechanisms. This hypothesis suggests that, within the genetic profiles of these dreadful TB strains, undiscovered and distinct gene variants exist that contribute significantly to antimicrobial resistance. These variants may encompass various genetic alterations, including single-nucleotide polymorphisms, insertions, deletions, and structural variations, collectively shaping the resistance patterns of the disease. The significance of this research extends beyond identification, as it lies in the potential of these variants as therapeutic targets. Such insights hold the promise of tailoring interventions and treatments specific to MDR and XDR TB types. The goal is to make substantial progress in effectively addressing the challenge posed by drug resistant TB strains, with the ultimate aim of improving global public health outcomes significantly.

1.6 Research Methodology

The methodology that has been adopted to carry out this research work is divided into four phases from data retrieval to mining significant patterns from the database and then from antimicrobial resistance gene identification to network analysis and hub genes identification.

1.6.1 Data Retrieval from NIAIDS TB Portal

- Data retrieval files, covering demographic details, DST results, and genomics (NCBI accession numbers).
- Acquisition of genomic data in fasta format from NCBI using python script.

1.6.2 Feature Selection and Data processing

- Feature selection and data preprocessing were carried out separately. Demographic data was utilized for explanatory data analysis, genomic data were employed for variant identification through ARIBA, and drug susceptibility data were analyzed for data mining purposes.
- Subseting the initial dataset by removing the cured cases, proceeding with drug resistant isolates.
- Removing the features which were not required for the analysis and deleting empty rows.
- Converting DST dataset in binary format.

1.6.3 Pattern Identification using Explanatory Data Analysis

- Writing a Python script for explanatory analysis using Matplotlib and Seaborn packages.

- Using the python script to create histograms from the demographic data.
- Analyzing gender distribution in relation to treatment outcome, type of resistance, and lineage, utilizing color palettes for clarity in Seaborn's hue parameter.
- Visualizing Frequent resistance types (XDR, MDR non XDR) in conjunction with gender, age, and disease outcome.

1.6.4 Drug Resistance Pattern Identification using Datamining Through Association Rule Mining

- Writing a python script for association rule mining employing aproiri algorithm.
- Selecting appropriate values for support, confidence and lift.
- Performing association rule mining on drug susceptibility dataset.
- Filtering the meaningless association rules.
- Classifying the rules with respect to type of resistance.

1.6.5 Identification of Novel AMR Genes in Multi and Extensively Drug-Resistant TB using ARIBA Tool

- Writing a python script to load raw fasta files to ARIBA software
- Aligning the sequences with *M.tuberculosis* reference genome.
- Using ARIBA to identify antimicrobial resistance genes.
- Analyzing tabular results provided by ARIBA containing gene names, variant types, novelty of the variants, and the effects of the variants on the gene sequence.

- Writing a python script to perform comparative analysis of genes with variants identified by ARIBA within each sample of XDR and MDR strains.
- Identifying common and unique genes with variants among XDR and MDR strains separately.

1.6.6 Functional Enrichment Analysis using String Database

- Using STRING database for functional enrichment analysis to identify over-represented functions or pathways associated with drug-resistant TB among the antimicrobial genes identified using ARIBA.
- Uploading the list of genes and select Mycobacterium TB H37Rv organism and perform the search.
- Selecting the candidate genes and optimizing analysis parameters to generate a graph representing the functional enrichment analysis results within the STRING platform.
- Downloading the graph and TSV (Tab-Separated Values) files containing the results of the functional enrichment analysis from STRING.

1.6.7 Hub Gene Identification Through Cytoscape

- Using cytoscape on the output from functional enrichment analysis to identify hub genes.
- Analyzing the identified hub genes which are highly connected and central within the network.
- Downloading the ranked list of the genes.

Chapter 2

Literature Review

2.1 Background

TB is one of the most fetal diseases around the world and was considered as the leading cause of death during 20^{th} century and still scientists and physicians are facing new challenges like antimicrobial resistance to combat this deadly infectious disease [35]. Pakistan is the fifth most affected country of MDR-TB. TB occurs in every part of the world but about 80 percent of the majority of TB-related deaths occur in countries with economies classified as low and middle income [1]. It is a Gram-positive bacterium and its genome comprises about 4.4 megabase pairs. *M.tuberculosis* is also an acid-fast organism which contains large amounts of mycolic acids within their cell walls contains around 4,000 genes and has a very high guanine + cytosine content that is reflected in the biased amino-acid content of the proteins [36]. Inactive TB is not vigorous, does not exhibit any symptoms and is resistible, while active TB shows symptoms and is profoundly irresistible. Such individuals are immune compromised due to various reasons including insufficient diet or comorbidity of HIV [37]. Lungs are affected in 90% of the cases and causes pneumonic TB. The patient develops various symptoms, for example, a cough for 14-21 days, bloody sputum, chest discomfort, weight loss, short breath, loss of appetite, fever at night [38].

2.2 Extra-Pulmonary TB (EPTB)

Extrapulmonary TB develop when the bacilli attacks organs other than lungs. Patients with HIV are at high risk of developing EPTB. EPTB constitutes about fifteen to twenty five percent of all the cases in susceptible patients. It includes TB of lymph nodes, cutaneous membranes, genitourinary tract infection, pericardial TB, TB in the joint and bones, radiation in the pleural cavity, TB in the larynx and the TB of meningitis [39]. The respiratory tract of individuals is contaminated and then the tubercle bacilli spreads through the lymphatic frame- work and then through circulation system to various organs. The advancement of tubercular sickness is partitioned into essential aspiratory TB, inactive TB contamination and dynamic TB illness, comprising tuberculo-meningitis, scrofulous TB, skin TB, cordis TB, pleurisy TB, urinary fundamental TB, stomach related foundational TB, skeletal TB, and so on [40, 41]

2.3 Transmission

The causative pathogen of TB, spreads by contaminating the environment. When infected person cough or sneeze. The microorganisms are pushed outside body, where these can be inhaled in by others. It never spreads by other physical contact like holding hands or usage of the same objects. The spread of the disease is elevated when these microbes unleased in the air can remain there for long it happens due to poor ventilation. It is possible that disease spread even if the carrier is long gone from the particular location the highly populated areas have high risk of the infection spread [42].

2.4 Pathogenesis

The dispersion of contaminated aerosols in the air derives the infection cycle of TB. even with presence of one to 10 bacilli can lead to disease transmission. As soon

as the bacilli enters the patient's lung, the alveolar macrophage cells engulf them, facilitating their invasion to the underlying epithelium. Here the granuloma is formed as the host immune system attempts to destroy the bacterial cells with the help of monocytes from surrounding blood vessels in host immune system [43]. The structure of granuloma comprises foamy macrophages, mononuclear phagocytes, and lymphocytes encapsulating infected macrophages and necrotic tissues debris. The formation is seemingly deceiving as the caseous core liquefies and cavitate and drains a large quantity of bacilli into the bronchi. The inflammation destroys lung tissue and result in, cough that, once again, contaminates the surrounding of the individual. This is one of the stages of TB [44, 45].

Infected macrophages use lymphatic system to travel the other extrapulmonary sites like lymph nodes, kidneys, epiphyses of the long bones even in the blood of an immunocompromised host such as HIV [40, 43].

The second stage is the growth stage where the bacilli continue to reproduce for three weeks, During the fourth week of onset, the infection enters the third stage known as the immune control stage. Now the bacilli progression and macrophages destruction are balanced [46]. In majority of cases the pathogens become dormant, they may reactivate after a while if the immune system gets compromised and enters the fourth stage which is the lung cavitation stage. The reactivated bacteria reproduce rapidly and develop a cavity in the tissue, where they are safe from the patient's immune cell. Then the bacteria rapidly spread throughout the tissues and patient begin to show the symptoms of active TB. The disease is extremely contagious at this phase [47]. Figure 2.1 illustrates the events initiating a complex interplay of immune responses and pathogens. Active infection/clearance, where macrophages and Th1 cells release TNF-α and IFN-γ to recruit immune cells, potentially clearing the infection or allowing bacilli multiplication.Active infection is followed by the formation of solid granulomas, composed of various immune cells, fibroblasts, and calcification, which Mtb manipulates for its survival by inducing the release of crucial chemokines and factors.The final stage is reactivation, where Mtb reactivates and exits the granuloma, typically under conditions of compromised immunity, such as poor nutrition or HIV coinfection, spreading to new

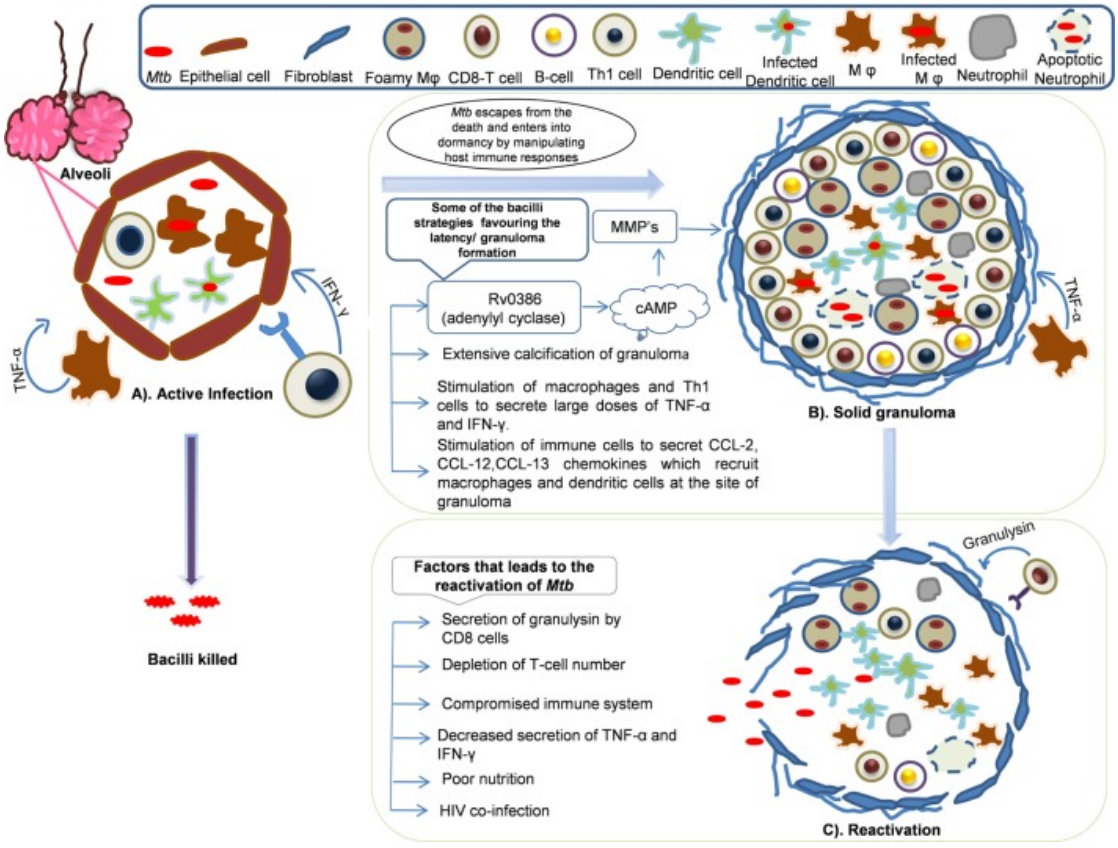


FIGURE 2.1: Overview of Pathogenesis

infection sites. This dynamic interaction between the host and Mtb shapes the course of TB infection. Patients with dynamic TB infection can be treated when they seek medicinal help. However, people with inactive TB can take prescription so they would not suffer from dynamic TB illness [48].

2.5 Symptoms

The signs and symptoms of TB refers to part of body it manifests. The infection grows slowly and gradually. The symptoms can show up after months or even years of the incident of infection. Latent TB do not develop any symptoms as compared to active TB and is apparently non-contagious. The most visible symptoms of active pulmonary TB include severe cough for at least 3 weeks, discomfort in the chest and spitting blood or sputum with cough leading to breathlessness. Some

other symptoms include weakness or fatigue, noticeable reduction in body weight, loss of appetite, chills, fever, and sweating [49].

The extra pulmonary TB has wide range of symptoms depending upon the site of infection. Whereas some common symptoms include swelling in glands or joints, fatigue, abdominal pain, constipation ,abnormal color of urine , headache, stiff neck a rash may appear on different parts of body [39].

Immediate medical consultancy is suggested if in case of prolonged cough, chest pain or any of the mentioned symptoms.

2.6 Latency

Inactive TB is where tests of individuals are positive for TB and have no clinical indications. It is thought that inactivity may contain a range of states, people free from the malady, to the untreated one, sub-clinical illness. *M.tuberculosis* can be contained inside granuloma for a considerable length of time. This capacity of *M.tuberculosis* to lie lethargic might be a transformative technique [50].

In majority of patients the infection stops here without showing any symptoms. In their lungs, the TB bacilli and macrophages that swallowed them build a round complex – with TB bacilli and infected macrophages in the middle and healthy macrophages surrounding them. Often TB bacilli also infect the surrounding lymph nodes which is called primary complex [51].

Some unlucky patients enter the stage four when the primary complex does not settle, and the bacteria is re-activated after a period of one to two years after the onset of infection. In this stage, the person is highly contagious because his or her sputum contains active TB bacteria. Reactivation of TB bacillus mostly takes place when the immune system is compromised, such as with HIV infection or malnutrition [46, 52].

2.7 Activation

Microorganisms' successful replication can overpower the immune system and break down granuloma barriers, leading to the release of *M.tuberculosis* into the lungs and resulting in illness. When the immune system is compromised, as in the case of an HIV infection, the risk of disease occurrence increases. [50].

Various factors contribute to the likelihood of *M. tuberculosis* activation. These factors include the individual's susceptibility and the infectiousness of a person with TB, determined by the quantity of tubercle bacilli they release into the air. Another critical factor is environmental conditions, encompassing factors such as the concentration of infectious aerosol particles, spatial arrangements, ventilation, air circulation, and more. Lastly, exposure plays a pivotal role, accounting for proximity, frequency, and duration of contact or exposure to the bacterium. [53]

2.8 Diagnosis of TB

The diagnostic methods have a vast range of depending upon the equipment and availability of assays at a particular medical facility. Initial diagnostic methods can be breath examination through stethoscope, susception for swollen lymph nodes and inquiring for the other symptoms. In case of suspicion the first diagnostic test is performed on the skin. A drug called tuberculin is injected in the epithelial tissues and the body's reaction to the drug detects the presence of Mycobacterium TB regardless of it is latent or active stage. Moreover, blood culture examination can be used to specify latent or active infection. Such blood tests are called interferon-gamma release assays or IGRAs. Another common diagnostic approach is getting a posterior-anterior chest radiograph or X-ray (CXR). This Xray is used to detects any kind of chest abnormality. The appearance of lesions of different dimensions and appearance can be observed anywhere in the lungs may indicate the presence of pulmonary TB infection. Actually, the chest radiograph is not the specific diagnostic approach for TB but may be used to rule out the possibility

of pulmonary TB in a person who has had a positive reaction to a TST or TB blood test and no symptoms of chest disease. In many cases sputum samples is also observed to confirm the presence of mycobacterium TB [54].

Diagnostic tests designed for drug-resistant TB aim to identify strains of TB bacteria that are resistant to commonly used antibiotics. These tests play a crucial role in assisting healthcare providers in determining the most effective treatment options for patients. Various methods are employed to diagnose drug-resistant TB include Drug Susceptibility Testing (DST), Genotypic methods utilizing molecular approaches like Line probe assay and XpertMTB/RIF. Drug Susceptibility Testing (DST)is a significant approach widely utilized method for evaluating the sensitivity or resistance of TB bacteria to different antibiotics. DST involves cultivating bacteria from patient samples and subjecting them to various drug exposures to observe their reactions. This process aids in tailoring treatment plans according to observed antibiotic resistance patterns [55].

Another approach is Genotype Testing, utilizing molecular tests to investigate the genetic material of TB bacteria and identify specific mutations associated with drug resistance. These tests offer rapid identification of resistance to drugs, providing valuable guidance for treatment decisions. Phenotypic Tests analyze bacterial growth under specific drug conditions. These tests provide insights into resistance levels and are particularly useful for detecting MDR-TB and XDR-TB [56].

Line Probe Assays (LPAs), operating at a molecular level, directly identify drug-resistant mutations in TB bacteria within patient samples. LPAs yield swift results, excelling particularly in identifying MDR-TB and rifampicin resistance. Molecular advancements have notably impacted TB diagnostics. Nucleic-acid amplification tests (NAATs), such as Line Probe Assays, have revolutionized the landscape with their high specificity and sensitivity [57].

Line Probe Assays were the first molecular tests recommended by the World Health Organization. These assays have significantly reduced the time required for diagnosing multidrug-resistant and rifampicin-resistant TB (MDR/RR-TB) compared

to culture testing. A significant step followed with the WHO's endorsement of the Xpert MTB/RIF assay in 2010. This, along with the Xpert Ultra assay, has markedly enhanced TB and RR-TB diagnosis compared to sputum smear microscopy, even at peripheral health system levels.The Xpert MTB/RIF Assay is a molecular test that concurrently detects TB and rifampicin resistance in a matter of hours. Rifampicin resistance often serves as a useful indicator of MDR-TB [56, 58].

For comprehensive insights, Whole Genome Sequencing takes the spotlight—an advanced technique that deciphers the complete genetic code of TB bacteria. This intricate method allows the identification of resistance mutations against a broad spectrum of drugs. It delivers a high degree of accuracy but requires specialized equipment and expertise [59].

These diagnostic tests play a critical role in managing TB, particularly drug-resistant forms. They ensure that patients receive suitable treatments to enhance outcomes and prevent the dissemination of resistant strains.

2.9 Treatment of TB and Drug Resistance

The treatment of TB depends on the effectiveness of drug regimen or vaccination. It is necessary to have an insight of the pathogen and immune response dynamics in order to understand the drug and vaccine efficacy.

However, the Bacillus Calmette Guerin (BCG) antibody is commonly used vaccine over the world. It is about 80 percent successful to shield youngsters from extreme type of TB meningitis which influences the cerebrum. BCG immunization likewise offers slight insurance in adults. The cure of tuberculous illness cannot be conceivable without an enhanced antibody [60].

More or less all of the antibiotic drugs of TB are effective while the bacteria are actively dividing. In this intensive phase of treatment, the drugs primarily kill quickly dividing bacteria, which soon results in rapid decrease of pathogens from

sputum, and the clinical symptoms disappears. The antibiotics used in this stage are called first line drugs (FLD). Nevertheless, for the eradication of stubborn or slow growing strains of *M.tuberculosis*, treatment enters the continuation phase which means the introduction of second line drugs [61].

2.9.1 First-Line Drugs (FLD)

The use of first line drugs (FLD) has resulted in the successful treatment of TB to some extent but, in some cases, it is perceived that these drugs unable to treat TB hence result in drug resistance due to several reasons [62]. The main FLDs are listed below:

2.9.1.1 Isoniazid

Isoniazid (INH) was discovered in 1952 as specific anti TB drug, since then it is being used as one of the main cure for the latent TB infection. It is a small molecule which is soluble in water hence is easily diffused in mycobacterium. It acts most effectively on frequently reproducing bacteria, but unfortunately, not so effective for non-dividing bacteria [63].

INH is a prodrug which is activated by mycobacterial catalase peroxidase encoded by katG.Then this activated INH suppresses two bacterial enzymes which are acyl protein carrier reductase which is encoded by inhA and acyl protein carrier kinase encoded by KasA [64]. These enzymes are responsible for mycolic acid synthesis which is a major component of mycobacterial cell wall. This way INH arrests bacterial cell wall formation and inhibits the disease [65].

Its resistant strain has been reported immediately after the discovery of the drug. The resistance is caused due to a number of factors but the major cause of resistance is mutations in the above mentioned genes i.e,KatG, inhA and KasA. Among these KatG mutation is the most severe mechanism and is also associated with its analog drug ethionamide (ETH) which is used as a second line drug [66].

Gene is not actually mutated but overexpressed due to various factors resulting in resistance to isoniazid [67].

The adverse effects of INH are associated with multiple neuropsychiatric symptoms like memory loss, hallucinations and even epilepsy because it crosses the blood brain barrier [66, 68].

2.9.1.2 Rifampicin

After isoniazid, rifampicin was second specified drug for TB, and it was reported in 1972 and is very effective. It basically targets the transcription of the bacteria by binding to the β-subunit of RNA polymerase encoded by the gene rpoB and plays leading role in transcription of bacilli [69]. Unlike isoniazid, it acts on slowly dividing bacteria beside rapidly dividing bacilli.

The most of *M. tuberculosis* samples from the patients who developed resistance against rifampicin had mutated gene rpoB which is codes for the β-subunit from RNA polymerase. This mutation results in structural changes in protein decreasing the affinity for drug, which hinders the bonding of drug and its target and consequence is drug resistance [70, 71].

2.9.1.3 Ethambutol

In 1966, the drug ethambutol was first presented as the cure of TB and still remains to be a part of first-line routine. Drug ethambutol is much dynamic against the effectively increasing bacilli, by targeting the arabinogalactan biosynthesis in dividing cell. The mycobacterial arabinosyl transferase catalyst is encoded by embCAB operon. Resistance from ethambutol is based on mutations in the gene embB [67]. Particularly the mutation in the multiple codons of embB gene is the cause of ethambutol resistance [72]. Ethambutol affects the formation of a structural unit of bacterial cell wall the arabinogalactan [73]. It hinders the polymerization of arabinogalactan and lipoarabinomannan in cell wall by accumulating D-arabinofuranosyl-P-decaprenol [74, 75].

2.9.1.4 Streptomycin

Streptomycin (SM), is an aminocyclitol anti-toxin, was used as the main medication for the treatment of TB. Streptomycin is aggressively dynamic against the moderate developing bacill. It binds with 16S ribosomal rRNA proteins; segments of the subunit bacterial 30s ribosomal subunit [76].

The two genes responsible for the resistance are; rpsL and rrs. S12 protein of ribosome and the 16S rRNA, are produced by rpsl and rrs genes separately, representing 60% 70% of streptomycin resistance [77, 78].

2.9.2 Second-Line Drugs

2.9.2.1 Fluoroquinolones (FQs)

These drugs are used against a wide range of bacterial infections including respiratory, gastrointestinal and urinary tracts, as well as sexually transmitted diseases. This category of drug includes ciprofloxacin, ofloxacin, levofloxacin, and moxifloxacin and are being used as second-line drugs in the treatment of TB.

The FQs mainly acts on DNA gyrase of *M. tuberculosis*. A type II topoisomerase protein which is made up of two subunits A and B which are encoded by gyrA and gyrB genes, respectively [79]. A small region of gyrA, called quinolone resistance-determining region (QRDR). Primary QRDR mutations responsible for of FQ resistance in *M. tuberculosis* are gyrB whereas gyrA are less frequent [80].

2.9.2.2 Second-line Injectable Agents

There are three injectable agents used in treatment of multidrug-resistant TB i.e., the cyclic polypeptide capreomycin (CAP), and the aminoglycosides amikacin (AMK) also a similar drug kanamycin (KAN). Both of the aminoglycosides (AMK, KAN) show elevated levels of cross-resistance between each other [81]. To counter

this situation the structurally different cyclic polypeptide CAP is a candidate that can be used as a substitute if resistance against AMK or KAN is observed [82].

AMK/KAN and CAP mainly acts on the protein synthesis of the bacterium and mutation in the 16S rRNA (rrs) cause resistance to these drugs [83]. There are different mutations nullifying the results of these drugs firstly the mutation C1402T primarily cause resistance against CAP and occasionally against KAN. Secondly the mutation G1484T results in high resistance against all three drugs [84].

Some other mutations in gene tlyA gene cause resistance against capreomycin. The gene tlyA codes for 2'-O-methyltransferase (TlyA) its mutation hinders mRNA tRNA translocation during protein synthesis [85].

2.9.2.3 Ethionamide/Prothionamide

Ethionamide (ETH, 2-ethylisonicotinamide) has been used against TB since 1956. There is another analogous drug prothionamide, both of these drugs are actually prodrugs, resembling isoniazid. These drugs are activated by a mono-oxygenase EtaA/EthA and share their target with INH. After entering the bacterium, ethionamide modifies itself. The drug is converted into 2-ethyl-4-aminopyridine after oxidation of its sulfo group by flavin monooxygenase. The transitional products formed before the synthesis of 2-ethyl-4-aminopyridine are lethal to the bacteria [86, 87].

2.9.2.4 P-Amino Salicylic Acid

The p-Amino salicylic acid (PAS) is among first antibiotics used against TB activity and was given in combination with isoniazid and streptomycin as FLD. Soon after the discovery of drugs such as rifampicin, it was placed in second line regimens. PAS is a worthwhile treatment of drug resistance TB, despite its limited benefits and high toxicity [88]

2.9.2.5 Cycloserine (CS)

Cycloserine is a medication used to treat TB, particularly in cases of drug-resistant TB like multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB). It works by inhibiting the growth of TB bacteria by affecting their cell wall synthesis. Often considered a second-line drug, Cycloserine is combined with other medications to improve treatment efficacy and reduce the risk of further drug resistance. Its use comes with potential side effects, particularly neurological, and is typically reserved for situations where other treatments have failed or when the TB bacteria are susceptible to it. Decisions about its use are made by healthcare professionals based on the patient's condition and medical history [89, 90].

2.10 Drug Resistance

Treatments of TB with anti-TB drugs have been used for decades and strains that are resistant to one or more of these antibiotics have been very well documented and studied. *M. tuberculosis* develops drug resistance when an anti-TB drug is used improperly, through inaccurate direction by health care providers, and patients stop medication before completion of treatment.

Patients might develop resistance to a single first-line anti-TB medication (isoniazid, rifampicin, ethambutol, or pyrazinamide), such resistance is called mono-resistant TB [91]. People also develop MDR-TB against a form of TB that does not respond to first-line anti-TB drugs INH and RIF. People with MDR-TB are treatable and curable by using second-line drugs that includes fluoroquinone and aminoglydins. Though, second-line treatment requires extensive chemotherapy for at least 2 years which is too expensive and toxic. In some cases, people also develop more severe drug resistance i.e., extensively drug-resistant TB (XDR-TB) which is a more serious form of MDR-TB. XDR-TB does not respond even to the most effective second-line anti-TB drugs, hence, leaving patients without any further treatment options [20]. Extensive drug resistance is divided into two cetagories

. Pre-extensively drug-resistant TB (Pre XDR TB) and XDR TB. Pre XDR TB arises from a strain that displays resistance to isoniazid, rifampicin, and either fluoroquinolones or injectable agents (amikacin, kanamycin, or capreomycin), but not both simultaneously. XDR TB is an exceptional type of MDR-TB that is resistant to isoniazid and rifampicin as well as to any fluoroquinolone and at least one out of the three injectable agents (amikacin, kanamycin, or capreomycin). Approximately 9% of the MDR-TB patients have extensively drug-resistant TB [91].

In 2023 according to WHO, the resistance to all first line anti-TB drugs (INH, RIF, Pyrazinamide (PZA) and EMB) was 26%, making MDR-TB as a public health crisis and a health security threat. Approximately, 450,000 new cases were reported by WHO to be resistance to the most effective first line anti-TB drug RIF, out of which 82% already had MDR-TB. It is estimated that only 55% of MDR-TB patients are currently successfully treated globally. *M. tuberculosis* strains comprises of seven lineages. Out of seven, four lineages are predominant in humans which includes lineage 1-4, i.e Indo-Oceanic, East Asian, East African–Indian and Euro-American [92]. Genomic studies have shown significant insights into the evolution of the *M. tuberculosis* and its resistance against anti-TB drugs. In 2021, 191,000 cases (from INH resistance strains) and 250,000 (from RIF resistance strains) cases of deaths, worldwide (WHO,2022). These resistant straints show the initial acquisition of INH resistance, followed by the resistance to RIF or EMB, then resistance to PZA and in the end, resistance to second line and third line drugs [93–95]. Recent studies report that the attainments of resistance, by spontaneous mutation have been estimated as 1 in 108 bacilli for RIF, 1 in 106 bacilli for INH, streptomycin and EMB, however, the rate of mutations in drug resistance strains is *M. tuberculosis* lineage-specific [96]. Among 7 lineages, the lineage 2 (for example Beijing lineage family) is highly associated with drug resistance in *M. tuberculosis* and has verified higher mutation rates in vitro studies [97, 98].

Many recent studies have shown information regarding the variation in the gene or genes encoding drug targets specific to resistant strains against the first line

drugs. Like other bacteria's, resistance in *M. tuberculosis* is acquired via vertical or horizontal gene transfer in fact it is mainly conferred by nucleotide variations [92, 99]. Thus, the basic mechanism for resistance in *M. tuberculosis* is the accumulation of point mutations (SNPs) in coding region of genes for drug targets and drug resistant disease arises through selection of mutants during insufficient treatment [100–103]. The frequently reported classical genes that are known to be linked with resistance against INH, RIF, EMB and streptomycin includes katG, inhA, rpoB, embB, rpsL, rrs and gidB, have shown mutation frequency of 70%, 10%, 95%, 70%, 6%, <10% and uncertain respectively in resistant *M. tuberculosis* strains *efpA, rpsL, katG, rpoB, blaC* [99] [104].

Where as in literature so far, there are nine known genes that have been identified in connection with resistance to the primary first-line TB drugs. These genes are associated with different drug resistances: *katG* and *inhA* are linked to resistance against isoniazid (INH), *aphC* and *kasA* are connected to INH resistance as well. rpoB is associated with resistance to rifampicin (RIF) [105, 106], while rpsL and rrs are genes linked to resistance against streptomycin (STR). Additionally, embB plays a role in resistance to ethambutol (EMB), and pncA is associated with resistance to pyrazinamide.These genes play pivotal roles in the development of drug resistance within Mycobacterium TB, the bacterium responsible for TB. Mutations or alterations in these genes can result in resistance to specific anti-TB drugs, complicating the treatment of TB infections [107]. An enhanced perceptive and knowledge of these resistant *M. tuberculosis* strains is urgently needed to direct recommendations for treatment of patients with first line drugs resistance.

2.11 Mechanisms of Drug Resistance

2.11.1 Antibiotic Modification or Degradation

Bacteria often become resistant to antibiotics by altering or breaking down the antibiotic molecules. This is especially common with antibiotics like aminoglycosides

(e.g., kanamycin, gentamycin, streptomycin), chloramphenicol, and β-lactams. For aminoglycosides, bacteria produce enzymes such as N-acetyl transferases (AAC), O-phosphotransferases (APH), and O-adenyltransferases (ANT) that modify the antibiotic, making it ineffective. These enzymes were first discovered in Streptomyces bacteria in the 1970s. An example of this mechanism is seen in Streptomyces griseus, where the enzyme streptomycin 6-phosphotransferase changes streptomycin into an inactive form, providing resistance [108]. Other antibiotics like bleomycin, tallysomycin, and chloramphenicol are also modified by specific enzymes to prevent their effectiveness [109].

2.11.2 Antibiotic Efflux

Another way bacteria resist antibiotics is by pumping them out of their cells using efflux pumps. This method is often used along with other resistance strategies. A well-studied example is found in Streptomyces peucetius, which produces the anticancer drugs daunorubicin and doxorubicin. These antibiotics are expelled from the bacteria by the DrrAB transporter system, an ABC transporter made up of the proteins DrrA and DrrB. This pump not only removes these specific drugs but can also expel various other drugs, similar to how the human P-glycoprotein pump works in cancer cells [110, 111]

2.11.3 Antibiotic Sequestration

Some bacteria resist antibiotics by sequestering, or trapping, the drug using specific binding proteins, preventing it from reaching its target. This mechanism is seen in the producers of the bleomycin family of antibiotics. In these bacteria, proteins such as TlmA, BlmA, and ZbmA bind to the antibiotics, either with or without metal, to stop them from working. These producers also have genes for ABC transporters in their antibiotic biosynthesis clusters, which likely help remove the sequestered antibiotics [112].

2.11.4 Target Modification/Bypass/Protection

Bacteria can also resist antibiotics by altering the target that the antibiotic aims to attack. For β-lactams, bacteria may produce more of the target proteins (penicillin-binding proteins or PBPs) or create PBPs that the antibiotic cannot easily bind to. For glycopeptides, resistance occurs when the bacteria change their cell wall precursors from D-Ala-D-Ala to D-Ala-D-Lac or D-Ala-D-Ser, reducing the antibiotic's binding ability. Other strategies include producing alternative versions of target proteins or enzymes, like DNA gyrase or RNA polymerase, which the antibiotic cannot effectively target [113]. Some bacteria also protect themselves by removing antibiotics from their targets, as seen with DrrC in Streptomyces peucetius, which removes daunorubicin and doxorubicin from DNA to allow normal cell function [114].

2.11.5 Pan Genomic Drug Resistance in Tuberculosis

Using comparative genomic analysis, particularly through pangenome construction, can help identify differences in how tuberculosis (TB) presents clinically. The pangenome includes all genes found in a species, split into a core genome (genes present in all strains) and an accessory genome [113].The accessory genome is crucial for phenotypic variation and evolution [115]. Though these genes are not essential for survival, they help bacteria adapt to different environments [116].

Several studies have examined the Mycobacterium tuberculosis pangenome. However, these studies have not compared the pangenome to differentiate between pulmonary tuberculosis (PTB) and extrapulmonary tuberculosis (EPTB) strains. This study aims to analyze the accessory genome in PTB and EPTB strains to find genomic markers linked to differences in disease presentation. Understanding these markers can reveal drug resistance mechanisms and aid in developing targeted treatments for TB [117].

2.12 Drug Resistance Data and Bioinformatics

As the amount of biological data is increasing, there is a dire need to analyze that data and drive conclusions. This data analysis can be either performed on DNA sequences, gene expression data, clinical data of some disease or any kind biological data.

In the world of bioinformatics, drug research data or drug resistance data can be handled in several ways. Some of the approaches applied on complex datasets, such as gene expression datasets, pan-genomes, metabolomics and biological pathways. Mostly genomic sequencing produces data for such applications. In terms of drug resistance this type of research questions is biased towards discovery and analysis of drug resistance genes, epistatic interactions related to antibiotic resistance, study of underlying regulatory mechanisms that results in resistance, or discovery of drug targets. The other applications deal with meaningful predictions of drug resistance based on clinical datasets. Such study is imitative from previous information of antimicrobial resistance mechanisms, and useful in finding generalize patterns and trends of genotype–phenotype relationships. Combinations of current molecular methods and powerful machine learning algorithms can be used for the understanding of antimicrobial resistance and improved clinically relevant predictions [118].

These predictions can manifest as patterns, figures, or facts, playing crucial roles in addressing various biological challenges, such as drug development or pathway design and analysis. [119].

2.13 Data Mining Techniques used in Bioinformatics

In the field of Data mining meaningful information is discovered, like associations, changes, anomalies, patterns and different structures, from outsized data which

stored information repository such as databases and data warehouses [120]. The arena of datamining has the abilities to determine unknown patterns and relations within the items in the biological data. In past few years, a rise in utilization of these techniques on clinical data is observed. The purpose is to identify convenient patterns which can be further applied for data analysis ultimately using it in decision making. In Data mining interesting information can be extracted from dataset by using one or more DM techniques which are clustering, classification, prediction, association learning etc. [121].

2.13.1 Classification

Classification is a categorization of data in a particular number of classes. The purpose is to recognize the categories or classes in which a fresh data can be placed. This process can be done automatically using algorithms. There are a number of classification algorithms some popular are logistic regression, Naïve Bayes, stochastic Gradient, K-Nearest Neighbours, Decision Tree, Random Forest, support vector machine [122].

In all classification algorithms, it is done in two steps, first training the dataset secondly testing. Training shapes the basic classification model on the using training data previously collected for producing classification rules. Most of the time, IF-THEN prediction rule is used which results in significantly useful abstraction. The accuracy of derived model pivot on the notch to which classification rules proved to be accurate, this is calculated by test data [123]. In terms of biology, we can say as "if Family History (for a particular disease) = yes & Consumption of Cholesterol = yes THEN Possibility of Disease = High", Such classification is beneficial.

Recently Pratik Sinha and colleagues developed a classifier which accurately identified acute respiratory distress syndrome (ARDS) from clinical dataset using a gradient-boosted machine algorithm [124].

2.13.2 Clustering

Unlike Classification, clustering does not construct classes instead large data sets are grouped in numerous of trivial subgroups known as clusters. This is done on the bases of similarities in dataset. Clustering algorithms determines assemblies of the facts e.g., items placed in same cluster could be more identical as compared to the ones in the other groups [125]. Famous clustering algorithms are K-Means Clustering, Mean-Shift Clustering, DBSCAN, Agglomerative Hierarchical Clustering, Gaussian Mixture Models. Clustering Algorithms have been used to scan the gene expression data [126].

In 2016 Kausar Noreen used K Means clustering and Support Vector Machines (SVM) algorithms on heart disease data successfully [127].

2.13.3 Association Rule Mining (ARM)

Association Rule Mining is an area of datamining which leads to the detection of associative relations or even correlations in a bunch of objects. There are many algorithms used for mining association rules, named as apriori, FP growth, IS algorithm, STEM algorithm and some variations of these. For biological data, these rules are proved reasonably convenient because they result in opportunity to conduct smart diagnosis and remove unimportant material and construct significant knowledge bases [120].

Basically, ARM is used to find the association rules that meets some predefined criteria. These conditions are minimum support, confidence, Lift and are decided by analyzing the nature of data. [128]. Association rule mining constitutes of solving two issues; finding all the frequent item sets and generating rules derived from frequent item sets.

TABLE 2.1: Brief overview of datamining usage in Health Science.

Area in medicine	**Data Mining Task**	**Algorithm Used**	**Ref.**
Pathology Data	Classification	Neural Networks	[129]
Coronary/Cardiac Diseases	Prediction and Classification	Decision Tree	[130]
		Algorithms	[131]
		Naïve Bayes	[132]
		Random Forest	[133]
Pulmonary diseases	Association Rule Mining	Aprori Artificial	[134]
	Classification	Neural Networks	[135]
Psychiatric diseases	Classification	ANN	[136]
	Prediction and Classification	BBN Bayesian Networks	[137]
Hepatic Illness	Association Rule mining	GP Growth	[138]
Dermatological Disease	Categorization and Classification	Decision Tree	[139]
		Artificial Neural Networks	[140]
		Naïve Baysian Algorithm	
Diabetes	Classification Association Rule	Support Vector Machine	[121]
	Mining	FP-growth	[141]
	Clustering	K-Means	[142]
Cancer/Breast Cancer	Classification Association Rule Mining	K-means, Apriori	[143–145]
	Naïve		
	Bayes		
Parkinson Disease	Clustering	K-means,	[146].
	classification	Naïve Baysian Algorithm	[147]

2.14 Related Work

2.14.1 Applications of Machine Learning Techniques on Healthcare Datasets

Several techniques have been applied to biological data, including drug resistance data to predict different aspects. Data mining methods, like association rule mining, have been applied on public health data in the past and effectively resulted in the discovery of interesting patterns. In [148], presented a framework which incorporated data from several clinics, discovered association rules, warehoused then for future use and provided the data is publicly available at an intuitive web interface.

In 2009 [149] discussed methods in data mining that can be applied in classification of data. She predicted the survival rate of a patient by selecting three data mining methods Decision tree, Naives Bayes and logistics regression on hospitals' data. Focusing on antibiotic resistance data in 2011, Mary Gerontini and fellows worked on predictions of associations in AMR and hospital borne infections. They presented an architecture in which they integrated data from multiple hospitals and discovers association rules stored in a data warehouse and used it as a source for extracting interesting and valid predictions by applying techniques such as regression and classification [150].

In 2017, Hayderpur and fellows worked on Nosocomial infections and antibiotic resistance Patterns from hospital in Iran. They collected data for antimicrobial resistance and analyzed it using SPSS and successfully identified useful patterns.

In 2019, Cazer analyzed MDR patterns in chicken-associated escherichia coli - linked multidrug-resistance dataset with association rule mining, also called market-basket analysis and they identified strong associations in antibiotics .In 2019 Konstantinos Vougas present a pipeline for screening AMR using association rule mining and predicted drug response [151].

2.14.2 Applications of Machine Learning Techniques on TB Datasets

Machine learning methods have been widely applied for timely predicting resistance of *M. tuberculosis* given a specific drug and identifying resistance markers [152]

In 2011 Tamer Uçar proposed the use of Sugeno-type "adaptive-network-based fuzzy inference system" (ANFIS) to predict the presence of TB on 667 different patient records consisting of clinical data [153].

In 2015 Ashwini performed a survey of machine learning approaches to detect TB in patients with or without HIV co-infection. In her work she discussed the main challenges in analysis and classification of clinical data. She finally classified the data into two classes first HIV with TB and second having HIV without traces of TB disease [154].

In 2018 Seelwan and krit, presented a Deep Convolutional Neural Network (DCNN) model, analyzing TB Chest X-ray (CXR) dataset of a population and compares it with non-TB CXR dataset of another population. The model forecasted that 36.51% of atypical radiographs in the CXR dataset were associated with TB [155].

In 2018, Srajan Kulshrestha applied various machine learning algorithms using antibiotic susceptibility test results as datasets. Patterns were identified using trends identified from results of dataset and were used to predict resistance towards various drugs [156]. Carly Bobak (2018) proposed a data analysis framework which directly integrated multiple expression array datasets in order to identify a more reliable gene signature for the diagnosis of TB. The method was successfully applied and diagnosed disease in 4 distinct datasets spanning a total of 1164 samples and 4 countries [157].

In 2019, Kamela from Belgium developed a tool in python 3 which give an alternative approach for attaining rifampicin-resistant TB diagnostic test results with whole genome sequencing instead of rapid diagnostic tests in laboratories [158].

Recently in 2020, Aytan-Aktug successfully predicted anti-microbial resistance using artificial neural networks on whole genome sequences of different bacterial specie [159]. Salma Jamal (2020) presented a computational framework that uses artificial intelligence (AI) based machine learning (ML) approaches such as naïve bayes, K nearest neighbor, support vector machine, and artificial neural network, for predicting resistance in the genes rpoB, inhA, katG, pncA, gyrA and gyrB for the drugs rifampicin, isoniazid, pyrazinamide and fluoroquinolones [160].

2.14.3 Use of Clinical Features to Determine Association Rules

During past few years clinical and demographic features are utilized through machine learning techniques for predicting patterns in order to understand different aspects of disease such as multi drug resistance. In Indonesia a pipeline was presented for medication selection process using rule mining clinical data on patients of 10 different diseases. The prescriptions of patients were used as clinical feature to recognize the relationship between the disease and the drugs advised by the physician. The analytical pipeline works in three phases first patient prescription data collection following the classification of the top 10 diseases using k-means algorithm and finally mining association rules using Apriori algorithm. The association rules provide the relation between the medicines and diseases. The medicines were prescribed on basis of support, confidence, and lift values [161]. Recently in 2021 Symptoms patterns of covid 19 patients were analyzed using association rule mining. Based on association rules, they concluded the repeatedly occurring were fever, cough, pneumonia, and sore throat. Whereas 1% of the patients exhibited severe symptoms, like septic shock, respiratory distress syndrome, and respiratory failure. The rules showed deviation in age and sex. Patients suffering from with chronic diseases had severe symptom rules such as, cardiovascular-related symptoms escorted by pneumonia, fever, and cough as consequents [162].

In another recent research association rule mining was used on the clinical features of 143 patients and established association patterns between three chronic

inflammatory diseases. A few patients affected by a combination of given 3 chronic inflammatory diseases which are: Type 2 diabetes mellitus (T2DM), Dyslipidemia (DLP) and Periodontitis (PD). This study considers almost 30 clinical features which are involved in these diseases. Then ARM was performed in order to derive consistent patterns among clinical features and diseases. Patients were divided into five groups based upon diabetic, dyslipidemic and periodontal conditions (including a healthy-control group).

At least 5 patients from each group were nominated to assess the gene expression analysis. ARM was performed on only CFs; and CFs+DEGs (Differentially expressed genes) to identify impactful associations. Then Gene expression validation was performed by running qPCR on Identified DEGs, specific to each group of patients. ARM proved to be an effective mining approach to analyze gene expression with the advantage of including patient's Clinical Features [163].

In Most of the research on drug resistant TB, considering the clinical features is an important for the treatment of disease. The clinical features being used commonly in drug resistant TB are weight loss, DST Profile, whether lungs are involved or not, age, gender, co-occurrence with HIV,smoking history and history of relapse. [164–166]. Recently in 2020, indian researchers associated clinical features and radiographic findings with drug resistant against the drug linezolid in 343 MDR TB patients. The symptoms they considered were weight loss, DST Profile against populat anti-TB drugs, Radiographs, and previous treatment history. They conclude that, DST is an important tool to identify linezolid resistance [167].

2.15 Gap Analysis

MDR and XDR TB is a crucial issue currently in management of MTB. According to a recent study, Asian countries such as Korea and the Philippines had significantly increased prevalence rates of approximately 600 TB cases per 100,000 persons annually in the past years while in Japan, incidence rates elevated due to a large number of people migrating to Japan from high burden TB areas for work

[168]. However, the highest incidence rate of MDR is found in China accounting for around 24% among treated cases [21]. A number of ML techniques like clustering, decision trees have been implemented to clinical datasets, even these techniques were used to investigate drug resistance but drug resistance of TB is not analyzed in order to extract resistance trends and patterns [148]. The majority of statistical techniques are limited because mostly the nature of MDR data happens to be non-Gaussian and sparse. Association rule mining (ARM) overcomes such limitations as it derives the interesting rules and patterns despite data distribution [169]. ARM have not been considered as an innovative tool for identifying patterns and trends of any kind of drug resistance [155].Bioinformatics can contribute in analyzing the patients using efficient machine learning algorithms to predict different patterns. Such Algorithms have provided positive results on clinical data of different diseases but TB. Hence applying machine learning techniques to clinical data pertaining to multi-drug-resistant TB is significant.

Chapter 3

Methodology

3.1 Introduction

In this study, dataset retrieved from TB portal was analyzed on the basis of multiple features to search for the existing patterns by utilizing pattern identification and datamining techniques. These patterns are further used to shortlist pathogenic isolates of XDR and MDR (non XDR) types of TB by employing WGS pipeline in order to identify novel antimicrobial resistant (AMR) genes that are specific to MDR and XDR strains of TB from United States National Institute of Allergy and Infectious Diseases (NIAID) TB Portals database. These gene/variants were annotated for their functional enrichment analysis to explore the role of drug resistance. After that the genes were further processed to identify the Hub genes using protein-protein interaction network that were significantly represented in some biological processes.

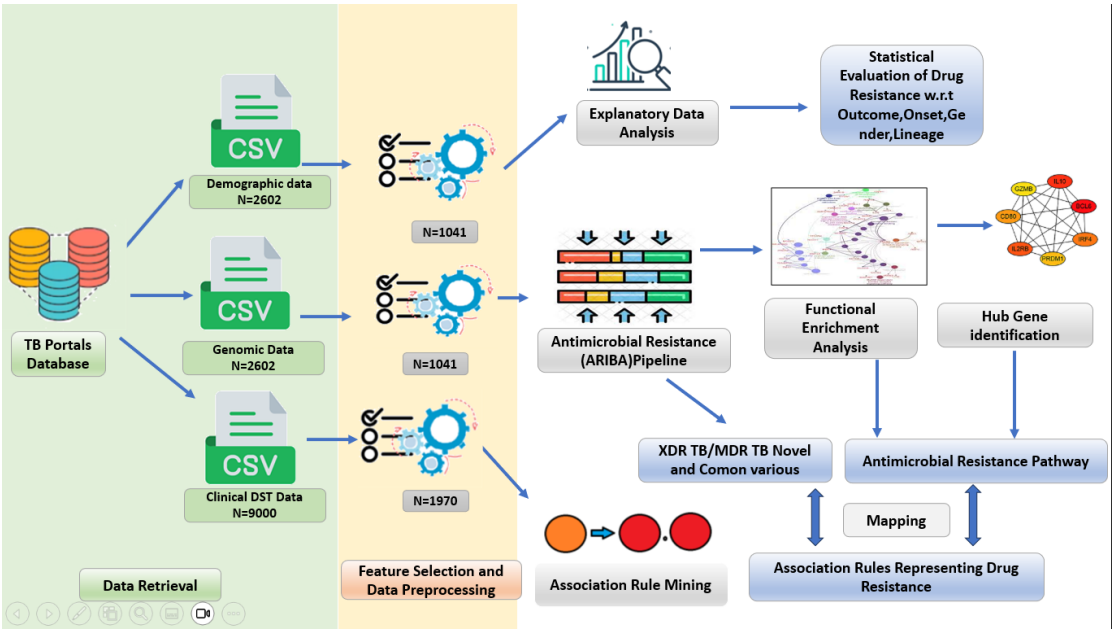


FIGURE 3.1: Overview of Research Methodology.

3.2 Tools and Equipment

3.2.1 Hardware specifications

The system used in the research was 11th Gen DELL with 8.00 GB RAM and Intel(R) core(TM) i5-1135G7 CPU. The processor specifications were 2.40GHz and 1.38 GHz.

3.2.2 Software

3.2.2.1 Windows Platform

Windows 10 Pro Version 22H2 with 64 bits Operating system, x64-based processor was used for experimental work.

3.2.2.2 Language Used In Project

The Pipeline used in this project was written in python 3.0. Specific scripts were written for a respective tasks. In Python the mlxtend library to perform association rule mining. With submodule 'frequent patterns.Moreover the functions 'apriori', and 'association rules' were in imported from the library.For Explanatory data Analysis the library 'matplotlib' was imported using pyplot module. Moreover 'seaborn library was also utilized.

3.2.3 Biological Databases

3.2.3.1 NIAID TB Portals Program

The data used in this project was acquired in collaboration with the NIAID TB Portals Program, (https://TBportals.niaid.nih.gov/). The permission to use the data was granted through MOU with the mutual consents and agreement signed by the scholar, supervisor and NAIB TB portals representative.The NIAID TB Portals Program is a multi-national collaboration for TB data sharing and analysis for advance TB research. The TB portals contain data of TB patient cases that have been contributed by multiple institutions from different clinical and research contexts. The TB Portals is an open-source web-based data repository containing a wide range of TB data and tools for analysis. There is no particular protocol for data collection. The information is submitted as part of usual practice by TB clinics, research studies, and clinical trials.

3.2.3.2 Sources of Data

The TB Portals has variety of data including socioeconomic/geographic, clinical (including MDR-TB data), laboratory, radiological, and genomic data collected from over 11000 TB cases from 18 sites in 15 countries throughout Eastern Europe, Asia, and Sub-Saharan Africa. (https://TBportals.niaid.nih.gov/). There are two types of data sources utilized by TB portals, one the TB Portals

Consortium cases and second the external cases submitted to TB portals. The TB Portals Consortium includes international allies in a direct collaboration with the TB Portals Program that assemble data and publicly shares with TB portals. The list of the institutions sharing data on TB portal is available in Appendix A.

3.2.3.3 NCBI GenBank

The NCBI GenBank, is a comprehensive and publicly accessible database that contains an extensive collection of genetic sequence data and supporting bibliographical and biological annotation maintained managed by the National Center for Biotechnology Information. Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references and a table of features (http://www.ncbi.nlm.nih.gov/collab/FT/index.html) that identifies coding regions and other sites of biological significance. GenBank serves as a valuable resource for researchers worldwide to access and analyze genetic information.

3.2.3.4 STRING

STRING is a biological database and web resource of known and predicted protein–protein interactions (https://string-db.org/).Several resources, including public literature, experimental data and the data acquired via computational prediction contributes to the information present in STRING database. It is publicly accessible, and it is consistently updated. The search engine also emphasizes functional enrichments in query lists of genes/proteins, using a variety of functional classification systems and databases such as GO, Pfam and KEGG.

3.2.3.5 GO (Gene Ontology)

The Gene Ontology (GO) is a knowledgebase globally renowned as the most extensive reservoir of insights regarding gene functions (http://geneontology.org/).

The Gene Ontology (GO) stands as a significant bioinformatics endeavor aimed at standardizing the depiction of attributes associated with genes and gene products across various species, provides annotations for genes and gene products, as well as gather and distribute annotation data and contains user-friendly tools to access all dimensions of the data present in the database /Additionally, enable the functional interpretation of experimental data through the use of Gene Ontology (GO), such as enrichment analysis.

3.2.3.6 PFAM

PFAM (Protein Families Database) is a database containing information about protein families and their annotations. It is a bioinformatics resource for identifying conserved regions and functional domains within protein sequences. PFAM classifies proteins into families based on shared sequence alignment and provides understanding of structural and functional characteristics of proteins. This data bases implements various algorithms including hidden Markov models.The PFAM database is constantly updated with new protein sequences and domain annotations as more genomic and proteomic data becomes available. The most recent version, PFAM 35.0, was released in November 2021 and contains 19,632 families. (http://pfam.xfam.org/)

3.2.3.7 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a prominent bioinformatics database offering insights into molecular interactions, pathways, and gene functions. KEGG is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances. (https://www.genome.jp/kegg/pathway.html). It provides pathway maps depicting relationships among genes, proteins, and metabolites.

KEGG aids in functional analysis, disease research, drug development, and evolutionary studies. Its tools enable sequence analysis and pathway mapping, while

also serving as an educational resource. KEGG is essential for understanding biological complexities and their applications in various fields.

3.2.4 Bioinformatics Tools

3.2.4.1 ARIBA

ARIBA (Antimicrobial Resistance Identification by Assembly) is a bioinformatics tool designed to detect antimicrobial resistance genes in bacterial genomes using whole-genome sequencing data. It utilizes a reference database of known resistance gene sequences for identification and employs a sensitive alignment-based approach to achieve high accuracy. ARIBA's automated workflow streamlines the process, and its reports offer insights into detected resistance genes, their locations, and related information. This tool aids in research and surveillance efforts, contributing to our understanding of antimicrobial resistance mechanisms and guiding treatment decisions. ARIBA is open-source and customizable, making it an asset for researchers and clinicians in the fight against antibiotic resistance.

3.2.4.2 Bowtie2

Bowtie2 is a software package commonly used for sequence alignment and sequence analysis in bioinformatics. It is a bioinformatics tool for aligning DNA sequencing data to a reference genome. It efficiently maps short DNA sequences to a target genome, supporting end-to-end and local alignments. It handles paired-end and mate-pair reads, reports multi-mapping, and balances sensitivity and specificity. Bowtie 2 is crucial for tasks like variant calling and gene expression analysis due to its speed and accuracy. Bowtie aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour.

3.2.4.3 SAMtools

SAM stands for "Sequence Alignment/Map." It's a file format used in bioinformatics to store alignment data from DNA/RNA sequencing reads to a reference genome. SAM files provide details about read alignments. SAMtools is a software package for working with SAM/BAM files, offering tools for conversion, sorting, indexing, visualization, and quality control. It's essential for processing high-throughput sequencing data. Within the ARIBA (Antimicrobial Resistance Identification By Assembly) workflow, SAM tools like SAMtools are utilized to process and analyze Sequence Alignment/Map (SAM) files. These tools filter reads, calculate coverage, and perform quality control, aiding in the accurate detection of antimicrobial resistance genes within bacterial genomes. SAM tools enhance the reliability and effectiveness of the ARIBA analysis pipeline.

3.2.4.4 CARD Database

The Comprehensive Antibiotic Resistance Database (CARD) (https://card.mcmaster.ca/) is a vital bioinformatics resource dedicated to antibiotic resistance in bacteria. It contains curated information about antibiotic resistance genes, associated proteins, and resistance mechanisms. Researchers use it to study and combat antibiotic resistance. CARD offers tools for searching and analyzing resistance data, follows standardized nomenclature, and is regularly updated by experts.

3.2.4.5 Cytoscape

Cytoscape is a open-source software platform used for visualizing, analyzing, and exploring complex networks, particularly in biology and systems biology. It offers tools for creating network graphs, conducting network analysis, integrating various data types, and benefits from a rich ecosystem of plugins. CytoHubba, is a plugin designed specifically for Cytoscape. It specializes in identifying and visualizing important hub nodes within biological networks. CytoHubba provides various

hub detection algorithms and allows users to customize their criteria for identifying these critical network components, making it a valuable tool for network analysis in biology.

3.3 Pattern Identification

The motivation of this study was to investigate the drug resistance patterns associated with *M. tuberculosis* and to determine the risk factors for XDR and MDR. Risk factor association was performed using explanatory data analysis techniques and for pattern identification of XDR and MDR datamining technique association rule mining was used.

3.3.1 Data Retrieval

TB portal provided different types of data including clinical, imaging, and bacterial genomic information, from both drug-sensitive and resistant cases. Data was retrieved in the form of csv files including data about demographic features, drug sensitivity test (DST) results, and genomic data. The DST data is clinical drug resistance data with 9000 entries, the sputum samples of patients were tested 27 drugs in vitro recorded in a single csv file. The samples which were declared resistant after DST testing were further subjected to genomic sequencing and the sequences were published in NCBI. Demographic data consisted of 19 features against 2602 fields. However, these features included the NCBI accession number of the pathogenic genomic sequence against each record. Python script was written to download the sequences for further analysis. Table 3.1 provides the details about drugs with the genomic targets according to the literature. The features and attributes used for analysis are mentioned in Annexure 2.

Two approaches to pattern identification were utilized: the first one was thorough explanatory data analysis while the second involved data mining. Two Distinct CSV files were employed for each method; for instance, demographic features were

TABLE 3.1: Anti TB drugs used for phenotypic drug susceptibility testing available in TB portal Data.

Sr. No.	**Drug Name**	**Type**	**Target Genes**	**Ref.**
1	Isoniazid (INH)	First Line Drug	*katG, inhA*	[66]
2	Rifampicin (RIF)	First Line Drug	*rpoB*	[71]
3	Streptomycin (STR)	First Line Drug	*rpsL, rrs*	[78]
4	Ethambutol (EMB)	First Line Drug	*embA, embB, embC*	[75]
5	Pyrazinamide (PZA)	First Line Drug	*pncA, rpsA, panD*	[170]
6	Ofloxacin (OFL)	Second Line Fluoroquinolones	*gyrA, gyrB*	[171]
7	Capreomycin (CAP)	Second Line Injectable Agents	*TlyA*	[172]
8	Amikacin (AMK)	Second Line Injectable Agents	*Rrs*	[173]
9	Kanamycin (KAN)	Second Line Injectable Agents	*Rrs*	[173]
10	Levofloxacin (LVX)	Second Line Fluoroquinolones	*gyrA, gyrB*	[174]
11	Moxifloxacin (MOX)	Second Line Fluoroquinolones	*gyrA, gyrB*	[175]
12	p-aminosalicylic acid (PAS)	Second Line Drug	*DHFR*	[101]
13	prothionamide (PTH)	Second Line Drug	*ethA, InhA*	[87]
14	Cycloserine (CS)	Second Line Drug	*Ddl*	[176]
15	Ethionamide (ETH)	Second Line Drug	*ethA, InhA*	[87]
16	Linezolid (LZD)	Second Line Drug	*RplC*	[177]
17	Fluoroquinolones (FQs)	Second Line Drug	*gyrA, gyrB*	[178]
18	Aminoglycosides injectable agents (AGs)	Second Line Drug	*whiB7*	[179]

employed in the EDA pipeline, while drug sensitivity data was used for the data mining approach. Feature selection and preprocessing was performed separately for each dataset.

3.3.2 Explanatory Data Analysis

3.3.2.1 Data Preprocessing and Feature Selection for Explanatory Data Analysis

The collected dataset contained a total number of 2,602 observations and 19 different features against each observation. Subsequently, 1,140 observations were initially subset by omitting those which showed sensitivity towards the drugs and the ones that were cured. Afterwards, the dataset was further subset according to the most common types of resistance (XDR, MDR non-XDR) found among all the samples. After this sub setting,1,040 observations were obtained that were next used for further analysis. Demographic Features selected for next step were gender, age of onset, type of resistance, outcome, and lineage. The remaining features were utilized for further genomic analysis as they contain the NCBI sequence accession information and snp information.

3.3.2.2 Identification of Multidrug Resistant Isolates with Respect to Age, Gender, and Outcome Through Explanatory Data Analysis

The TB dataset, which includes various features such as gender, age of onset, type of resistance, outcome, and lineage, was analyzed using data analysis and visualization techniques. Explanatory data analysis was carried out to summarize the main characteristics of dataset to understand and explore the relationship of clinical descriptors with XDR and MDR. For this analysis EDA was performed using a customized python script which utilized matlib and seaborn packages which are the basic tool for explanatory data analysis. The script is available in annexure 3. Each descriptor was utilized separately using data visualization technique

available in seaborn. To achieve this, each feature was visualized separately using appropriate data visualization methods For instance, age distribution was elucidated by using optimal binning while visualizing its patterns through histogram [180]. Subsequently, some patterns and trends were obtained by the set of combinations, involving any two distinct features. Depending on the observed patterns, age of onset was compared with outcome of the disease, type of resistance, and lineage by using box plots to determine which age group has majorly faced the worst outcome, which type of bacterial resistance was responsible, and how lineage was being influenced in various age groups respectively. Similarly, gender distribution was compared with the treatment outcome, type of resistance, and lineage by utilizing the parameter hue, which uses color palette to visualize the recognized patterns with more clarity while comparing two or more than two features at the same time [181]. Moreover, through bar charts, the type of resistance was first compared with the disease outcome to understand how resistance of the bacteria influences the outcome of disease followed by its comparison with the lineage to analyze which lineage is significantly being affected by the bacterial resistance. Furthermore, most frequent types of resistance (XDR, MDR non XDR) were then visualized in combination with gender, age, and outcome of the disease.

3.3.3 Drug Pattern Identification using Datamining

3.3.3.1 Data Preprocessing and Feature Selection for Data Mining

The raw data sheet contained almost 9000 records and 127 columns. The multiple sputum samples were tested against different 18 TB drugs. Since data present in TB portals is sourced from diverse laboratories, each adhering to specific protocols for antimicrobial testing.These may include Bactec, Lowenstein–Jensen (LE), Line-Probe Assay (LPA), LPA-Hain, and GeneXpert. Consequently, variations emerge in drug names and dosages due to diverse lab practices. However the table follow a standardized structure, evaluating 18 drugs across all five DST test typed. It is noteworthy that not all columns contained data. The table contained

3 types of values, 'R', 'S', 'NULL. Drug resistant was denoted with R, sensitivity was denoted with 'S' and undocumented data was denoted as 'NULL'. Data was preprocessed into a binary dataset. In order to do so R was replace with 1 and both other values were replaced with '0'. Further duplications were removed and empty cells were deleted. The binary data was organized as items and transactions, with one transaction per row and one item per column. The preprocessing was performed through python script available in Appendix C and further verified manually.

3.3.3.2 Pattern Identification of Multidrug Resistant Isolates Through Data Mining

The approach of data mining used for pattern identification was association rule mining. This step was carried out to find out most frequent patterns of drug resistance in terms of drug. Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases Association rules can expose biologically significant associations between diverse components such as drugs and drug target, genes and gene expression or drug's resistance with other factors. An association rule is represented in the form [L.H.S] => [R.H.S] where [L.H.S] and [R.H.S] are actually disjoint items sets, the [R.H.S] set has chances to occur whenever the [L.H.S] set occurs. There are two main steps for pruning association rules:

- Finding Frequent item sets
- Association Rule generation

3.3.3.3 Quality Measures

Quality measures in association rule mining are used to assess the significance and reliability of discovered association rules. These measures help determine

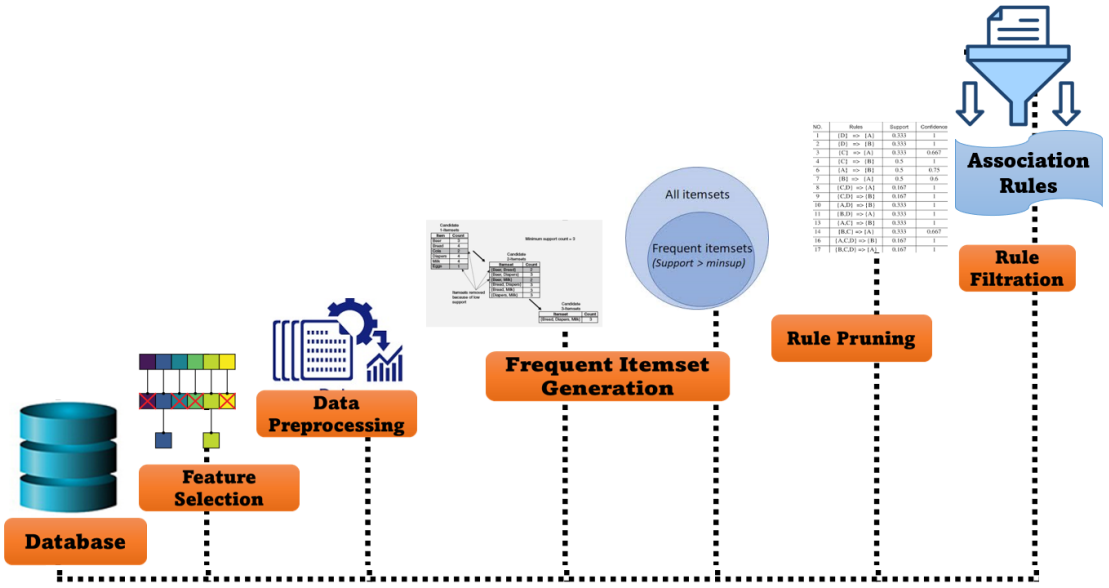


FIGURE 3.2: Association Rule Mining Methodology.

which rules are interesting, meaningful, and worth further investigation. The Two important quality measures considered in this study are support and confidence.

3.3.3.4 Support and Support Count

Support Count is the frequency of transactions that contains a particular Item set. Support measures the proportion of transactions that contain a particular item or rule in the dataset. It indicates how many transactions contain both the antecedent and consequent of the rule. Support calculation in association rule mining is essential to discover frequent itemsets in the data. It is the quality measure that plays a crucial role to filter out infrequent or uninteresting itemsets, prune less significant rules, prioritize significant associations, and provide insights into data patterns. Support is an important quality measure that warrants the significant discovery of associations in large datasets. High support suggests that the rule is frequent in the dataset. Support is calculated as follows:

Support(X) = (Number of transactions containing X) / (Total number of transactions)

Here 'X' is an itemset,

'Number of transactions containing X' means the count of transactions in which

the itemset X is observed. 'Total number of transactions' means overall number of transactions in your dataset.

3.3.3.5 Confidence

Confidence measures the strength of association between the antecedent and consequent of a rule. It is calculated as the ratio of the support of the combined itemset (antecedent and consequent) to the support of the antecedent alone. Confidence in association rule mining quantifies the strength of the association between items in a rule. It calculates the frequency of the consequent item appears when the antecedent is present. High confidence indicates a strong relationship between these items, making the rule more reliable and meaningful. This is also called conditional probability. The pre-defined minimum confidence (minconf) is used to select reliable rules from all possible rules. confidence is a measure of the strength of an association between two itemsets, X and Y, in a rule of the form "if X, then Y." It quantifies the probability that itemset Y will be present in a transaction given that itemset X is present in the same transaction. The formula for calculating confidence is as follows:
Confidence(X => Y) = Support(X ∪ Y) / Support(X)
Here, Support(X ∪ Y) means the support of the combined itemset (X ∪ Y), meaning the number of transactions where both X and Y are present.

3.3.3.6 Lift

Lift is another important measure in association rule mining that assesses the strength of association between items while taking into account the expected frequency of their co-occurrence. The formula for calculating lift is as follows:
Lift(X => Y)=(Support(X)×Support(Y))/Support(X ∪ Y)
Lift values greater than 1 indicate a positive association, suggesting that the presence of one item is likely to increase the presence of the other item, meaning they are associated or correlated. Lift values equal to 1 indicate independence, meaning

that the presence of one item does not affect the presence of the other item. Lift values less than 1 indicate a negative association, suggesting that the presence of one item is likely to decrease the presence of the other item, meaning they are mutually exclusive. Lift is a useful metric because it not only measures association but also provides information about whether the association is significant or just due to chance.

3.3.3.7 Association Rule

An association rule is a pattern or relationship discovered in a dataset that shows the statistical association between items. Association rules are typically expressed in the form of "if A, then B" .Where A and B are sets of items or attributes. The strength of the association is measured by quality measures like support and confidence. Association rule mining in drug resistance data is essential for identifying trends of co-resistance patterns of XDR and MDR in patients. A dataset can contain $3^n - 2^n + 1 + 1$ potential rules where n is the is the total number of transactions in the dataset.

3.3.3.8 Frequent Itemset Generation

If the support of an item set is equal to or more than predefined support value then that particular item set is frequent item set otherwise infrequent item set. In context to drug resistant data.An itemset is a group of 0 or more items (i.e. drugs). If all the items in the itemset are present in a transaction then we can say the transaction has that itemset. The number of possible item sets, excluding the null set of zero items, is 2^n - 1, where n is the number of items in a dataset.

3.3.3.9 Rule Generation

The Apriori algorithm was employed to identify association rules using a custom written python script (see Appendix C) . This algorithm is capable of discovering frequent item sets by choosing candidate itemsets on the basis of minimum

support threshold. The support of an itemset should be rather less than or equal to the support of its subset.To calculate association rules using the Apriori algorithm in Python, the mlxtend package was utilized for Apriori implementation. Additionally, the pandas package played a role in efficient data manipulation, especially when working with data in DataFrame format. The process involved the preparation of the dataset, specification of a minimum support threshold to find frequent item sets using Apriori, and the subsequent generation of association rules based on these frequent item sets. The algorithm considers the smallest itemsets (with one item) and eliminate the set which does not meet the minimum support requirement. Consequently, all candidate item sets with infrequent items were eliminated because they cannot not meet the minimum support. The algorithm generates all possible two-item item sets and removes the candidates with their support less than or equal to minimum criteria. Similarly, the supports of the rest two-item itemsets were calculated and compared to the min support criteria. The algorithm repeated this procedure to the point where all item sets of a assumed size were discovered to be infrequent or the algorithm reaches the largest candidate item set. In this way aproiri successfully recognized the frequent itemsets without calculating the support of each possible itemset.

Once the frequent item sets had been established, association rules were generated from these sets. These rules were formulated under a condition that ensured the support of each rule exceeded or equaled the minimum support threshold. Following this step, the frequent item sets are partitioned into two distinct and non-overlapping subsets: the antecedent and the consequent. This partitioning effectively outlines the connections and dependencies among the items within the rule. This method allows for the extraction of significant associations within the data, thereby aiding in the comprehension of drug resistance patterns and their implications. Once the frequent item sets had been established, association rules were generated from these sets. These rules were formulated under a condition that ensured the support of each rule exceeded or equaled the minimum support threshold. Following this step, the frequent item sets were partitioned into two distinct and non-overlapping subsets: the antecedent and the consequent. This

partitioning effectively outlined the connections and dependencies among the items within each rule. This approach facilitated the extraction of significant associations within the data, contributing to a better understanding of past drug resistance patterns and their implications.

3.4 Identification of Novel AMR Genes in Multi and Extensively Drug-Resistant TB using ARIBA Tool

Antimicrobial Resistance Identification by Assembly (ARIBA) is a tool used to detect Antimicrobial Resistance (AMR) genes by analyzing paired read data. It takes raw sequencing FASTA files as input and aligns the reads to a reference genome.

These raw sequences were downloaded from NCBI. For downloading the sequences, the SRA number provided in NAIDS data was used for accession number. These 1041 raw FASTA files were retrieved and used as the input for ARIBA software and was loaded through a python script. The reference genome used was the *M. tuberculosis* reference genome obtained from the NCBI Genomes database. Once the reads were aligned, ARIBA identified variants for each gene in the sample. It provided tabular results that include information such as gene names, variant types, novelty of the variants, and the effects of the variants on the gene sequence.

The analysis aimed to investigate the presence of novel coding gene variants that could potentially be linked to antimicrobial resistance in two types of *M. tuberculosis* strains, namely extensively XDR and MDR strains. Custom Python scripts (Appendix C) were written to perform comparative analysis of genes with variants identified by the ARIBA tool within each sample of XDR and MDR strains.(Fig.

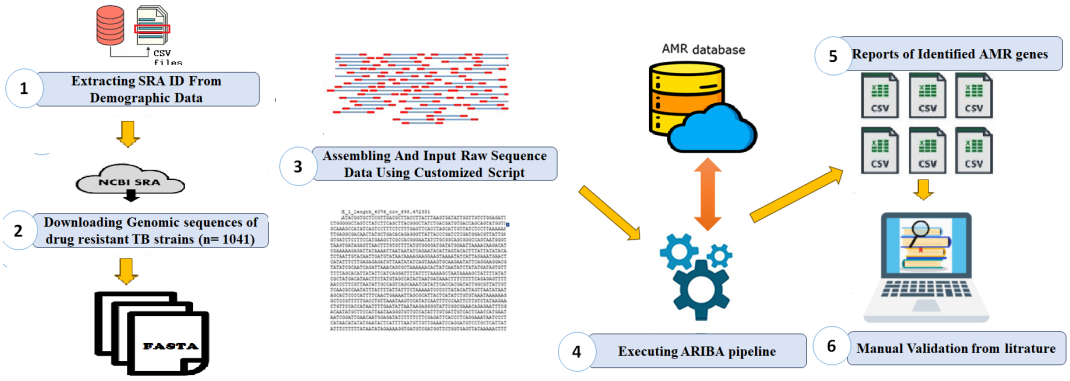


FIGURE 3.3: AMR genes identification using ARIBA Pipeline

3.3)This task involved reading each CSV report generated by ARIBA and comparing among each of the isolate to identify the common genes that have variants among XDR and MDR strains, separately. Moreover, through this process, common novel variants shared by both XDR and MDR strains were also identified. Furthermore, Venny tool was employed to identify unique genes that were exclusive to either XDR or MDR strains, based on the common novel variants previously identified.

These unique genes represented genetic variations that were specific to either XDR or MDR strains. The common novel variants as well as the unique variants specific to each strain type were then selected for functional enrichment analysis. Ariba aligned query sequences based on homology and similarity, regardless of the bacterial species. Following manual inspection and a literature review, certain genes were identified that were either unrelated to *M. tuberculosis* or not typically associated with resistance mechanisms. After applying filters, the results were categorized into three groups of genes: 1) Unique novel variants specific to XDR TB and MDR TB, (??) 2) Common novel variants shared by XDR TB and MDR TB, (4.5)) and 3) Novel unidentified variants that have not been previously reported in the context of TB drug resistance.

3.5 Functional Enrichment Analysis

The MDR and XDR novel variants were classified into two categories: common and unique. Both the unique and common variant genes of MDR and XDR were subjected to functional enrichment analysis. To perform this analysis, the STRING database was utilized, which provides a comprehensive list of genes involved in various pathways. String provides a number of gene sets which can be used for enrichment analysis. Enrichment analysis is used to characterize a gene list by looking for classes of gene representing functions that are overrepresented on the gene list to associate with the drug resistant TB. String has a user-friendly interface. First of all the candidate genes/proteins were searched by organism name, "Mycobacterium TB H37Rv" was chosen from the provided list. A graph was generated after further optimization of the parameters provided in string functional enrichment analysis. The TSV files were downloaded and analyzed manually. For further analysis and insights the result was subjected to cytoscape.

3.6 Hub Gene identification

Through the utilization of STRING analysis, an examination of the common genes shared by MDR and XDR strains revealed distinct clusters of interactions. Cytoscape, a powerful tool in systems biology and network analysis, was used to identify hub genes. The input was the TSV files imported from functional enrichment analysis conducted on the list of MDR and XDR TB genes, revealing the biological processes, pathways, or functions significantly linked to these genes. Subsequently, within Cytoscape, a network is constructed where each gene is represented as a node, and edges symbolize known interactions or relationships between these genes. The CytoHubba plugin for Cytoscape was employed to identify hub genes within the network. Hub genes are typically genes that are highly connected or central within the network.. The network from string was subjected to cytoscape and the degree centrality algorithm was selected to identify the hub genes which was based on the number of connections they possess. CytoHubba

provided a ranked list of these crucial genes, which was be further visualized within the network, making them distinguishable from other nodes. The direct and indirect interaction of gene was revealed as the on the basis of placement of genes in a specific cluster. Furthermore, a comprehensive network of interactions was observed among all other genes, forming an intricate map. Notably, each gene exhibited the capability to receive and transmit signals while interacting with other proteins. These hub genes are potentially fundamental in the context of MDR and XDR TB, acting as crucial regulators or central players within the biological processes associated with drug resistance in TB.

Chapter 4

Results and Discussion

4.1 Identification of Multidrug Resistant Isolates with Respect to Age, Gender, and Outcome

In order to identify drug resistant isolates of TB, various demographic features present in the dataset (2,602 observations) were analyzed and shortlisted by using pattern recognition techniques. The patient with treatment outcome 'cured' were omitted as they were sensitive to the treatment, remaining 1041 observations were subjected to explanatory data analysis. Type of resistance which refers to the resistance exhibited by TB strain, age of onset which corresponds to the age at which TB was diagnosed, gender of the patient, and outcome of the patient's treatment were the selected features as they were identified with some prominent patterns and regularities in the data. In order to evaluate the most prevalent type of drug resistance among all the types of resistance present in the dataset(pre-XDR, XDR, sensitive, mono DR, poly DR, and MDR non XDR), a bar graph was plotted between the count of patients on y axis and x axis represented the type of resistance.(Fig. 4.1)Among 1041, 741 were XDR,48 were mono DR, 4 were pre-XDR, 311 were XDR and 38 were poly DR.The graph showed that most common and highest occurring type of resistance is MDR (71%) followed by XDR(29.8%), mono DR (4.6%), poly DR(3.6%) and pre-XDR(0.3%) . In order to influence on

drug resistance type on outcome of the disease (failure, lost to follow up, still on treatment, completed, and unknown), another bar chart was plotted to visualized bacterial resistance type with respect to all possible outcomes of the disease mentioned in database shown in (Fig. 4.2). It was noted that all corresponding outcomes were comparatively higher for MDR then XDR. Subsequently, isolates with XDR and MDR non XDR were separately plotted with outcome of the disease to visualize and analyze outcomes associated with these types of resistance by bar chart plot (Fig. 4.3). After final subsetting of the dataset, it was observed that the outcomes were Death (MDR 110/741 - 14.86%, XDR 100/311 - 32.16%), Treatment failure (MDR 150/741 - 20.24%, XDR 77/311 - 24.76%), lost to follow up (MDR 178/741 - 24.00%, XDR 30/311 - 9.65%), still on treatment (MDR 96/741 - 12.97%, XDR 47/311 - 15.11%), completed (MDR 182/741 - 24.56%, XDR 41/311 - 13.20%), and unknown (MDR 13/741 - 1.75%, XDR 16/311 - 5.14%). Analysis was also conducted to examine the distribution of resistance types based on gender. It was revealed from the bar chart that males are significantly being influenced by both types of resistance, as compared to females. XDR TB is present in around 220 males out of 311 while the count of females is less than 100 out of 311 observations.

Around 570 out of 741 males were found with MDR non-XDR TB, whereas only 160 out of 741 females possessed MDR non XDR TB.(Fig. 4.4) How ever male and female resistance type was analyzed separately which showed that over all resistance type ratio remains same for each gender. To achieve this, the initial dataset, comprising 2602 observations, was utilized. The dataset was then divided into two subsets: one for males and one for females, allowing for separate analysis of resistance patterns in each gender.Out of 2602, 736 (28%)were females that out of which 141 were sensitive and did not faced the worst outcomes, so these isolates were omitted leaving 594 observation. Among these 594 observations 351 (59%) were MDR, 175(29%) were XDR, 41(6.8%) were Mono DR,24(4%) were Poly DR and 3 (0.5%)were Pre-XDR .Total males were 1868 (72%).Similarly after excluding the cured observation which were 337(18%),Total 1531 observation were carried for futher analysis .Among these 1531 observations 998 (65%) were MDR, 391(25%)

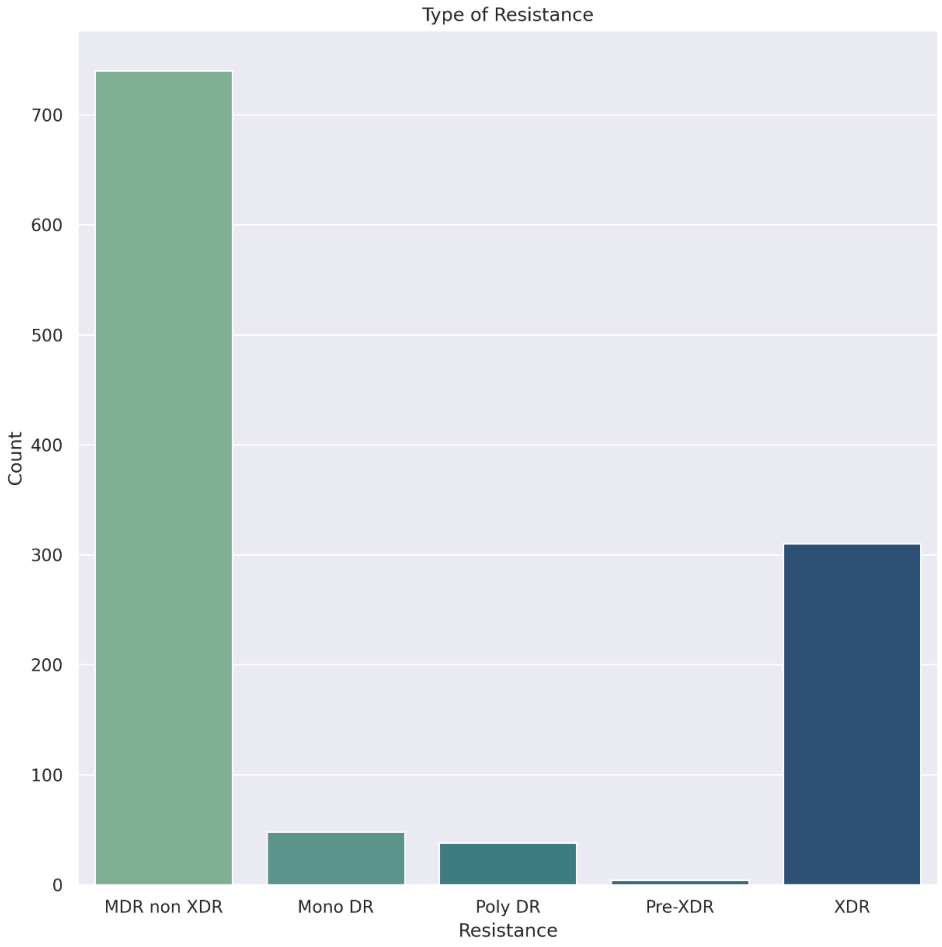


FIGURE 4.1: Bar chart demonstrating types of resistance against the count of patients after subsetting.

were XDR, 93(6%) were Mono DR,56(3.6%) were Poly DR and 2 (0.13%)were Pre-XDR .The percentage ratios of the resistance type remained almost the same. So It was concluded that gender has no specific impact on drug resistance.(Fig. 4.5)

Additionally, the patient's age of onset was then compared with the resistance of bacteria by visualizing through a comparative box plot (Fig. 4.6). This was done to analyze the onset age distribution among the isolates of the patient having MDR non XDR and XDR as the types of resistance. Age around 40 is the median, corresponding to both MDR non XDR and XDR TB. Moreover, the age of patients

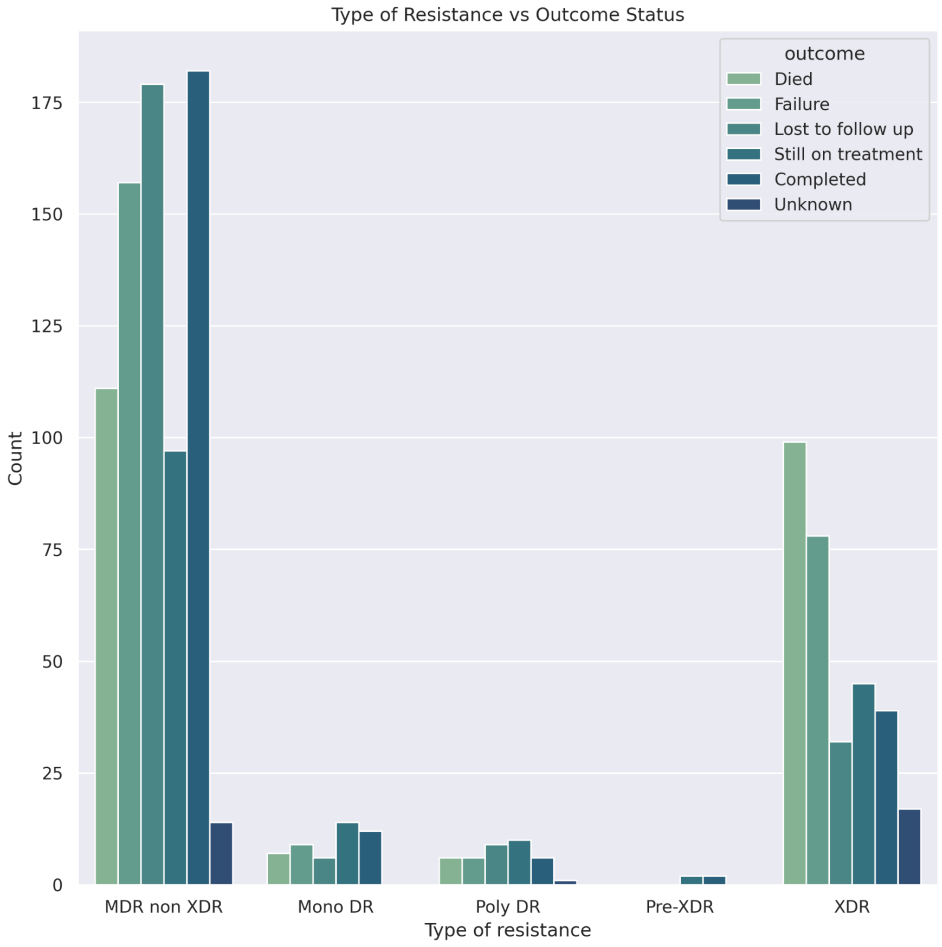


FIGURE 4.2: Bar chart demonstrating types of resistance against the disease outcome.

with XDR TB ranges from 18-75 years while the age for MDR non XDR TB patients ranges from 13-82 years. Majorly, both types of resistance range from ages between 32-52 years as demonstrated by the box plot. There was no significant difference in onset of resistance with respect to the age that indicates that age has no relation with the onset of any type of resistance.

In order to assess the type of resistance with respect to lineage another bar graph was plotted depicting that Beijing lineage following H3 and T1 showed higher prevalence of MDR as compared to other reported lineages. Bar graph was plotted that illustrates the relationship between different lineages of *M. tuberculosis*

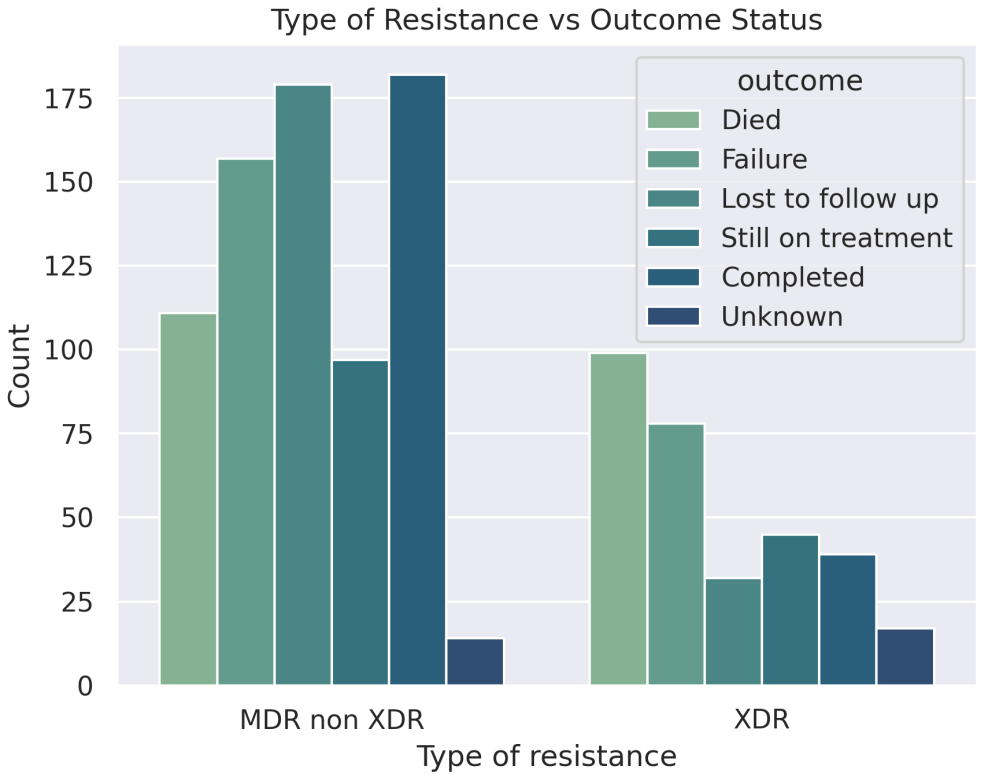


FIGURE 4.3: Bar chart demonstrating types of resistance against the disease outcome after subsetting.

and the types of drug resistance.(Fig. 4.7) While working with the DR-TB data, generation these bar plots were fundamental. The visual representation of the relationships between patient characteristics, drug resistance types, treatment outcomes, age, and TB lineage, provided valuable insights into DR-TB dynamics. These insights are essential for tailoring treatments, understanding treatment outcomes, identifying potential risk factors, and guiding public health strategies. The analysis revealed that XDR and MDR non-XDR were the most prevalent types of TB resistance and were associated with unfavorable outcomes. Specifically, significant number of patients with MDR non-XDR TB experienced death or treatment failure. Age of onset analysis through box plots demonstrated that the median age for both types of resistance was around 40, with the age range varying slightly between XDR and MDR non-XDR TB cases. In a study conducted in Sudan, the 25-44 and 45-64 age groups were more likely to be infected with MDR-TB than the other age groups (18-24 years and 65+ years). A case-control study conducted

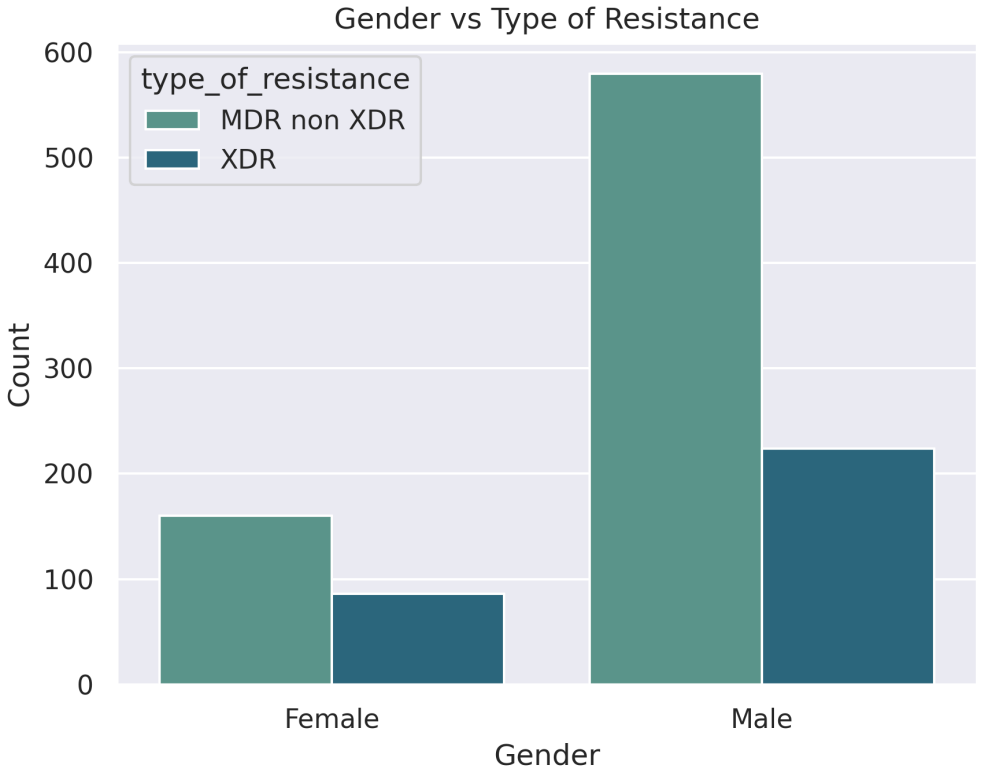


FIGURE 4.4: Bar chart illustrating gender distribution against the types of resistance (MDR non XDR, XDR) after subsetting.

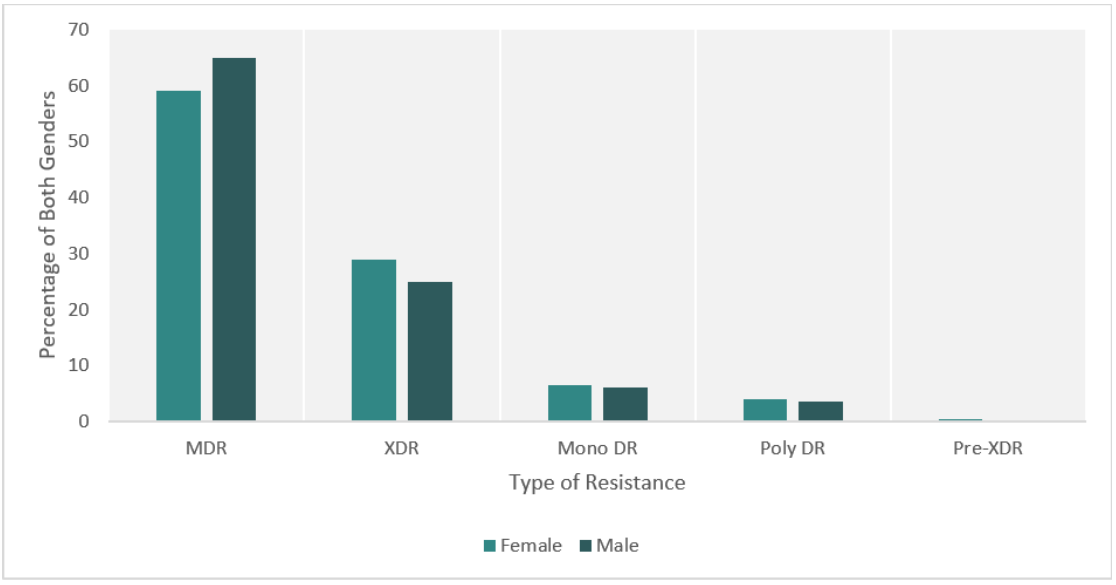


FIGURE 4.5: Box plot illustrating types of resistance (MDR non XDR, XDR) against the gender.

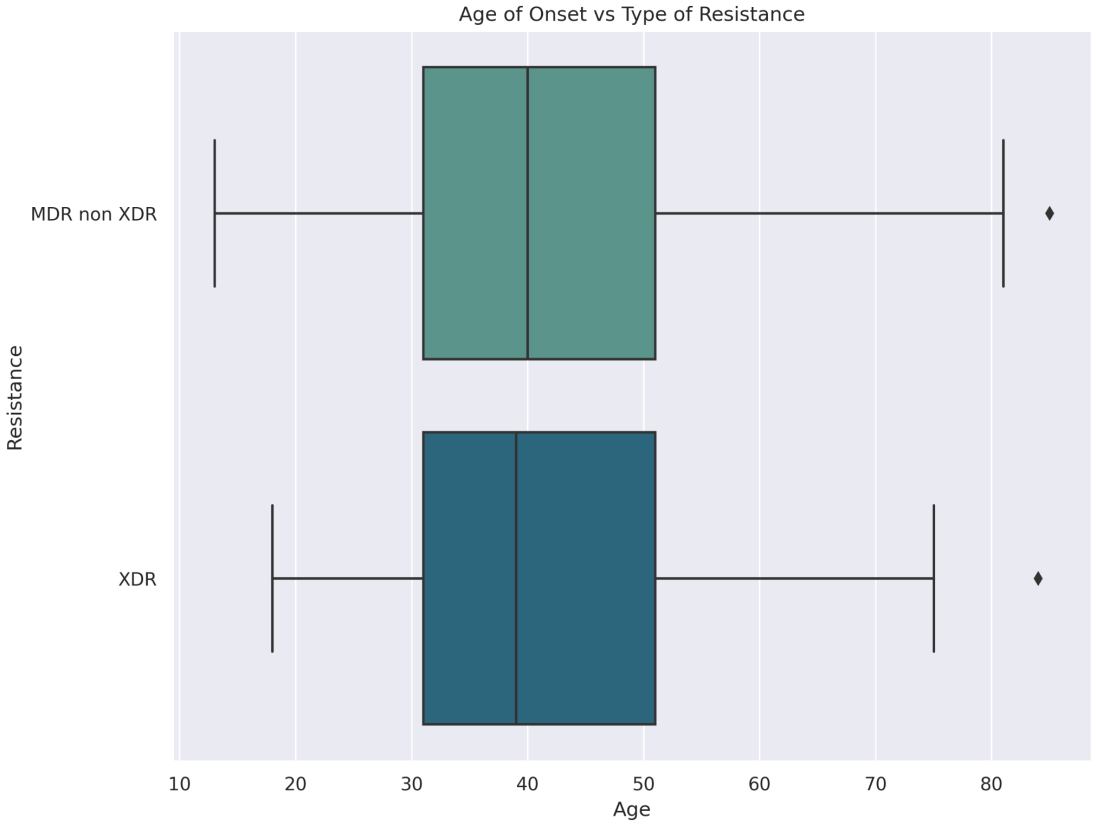


FIGURE 4.6: Box plot illustrating types of resistance (MDR non XDR, XDR) against the age of onset.

in Bangladesh confirmed this finding in the 25-44 age group, among whom MDR-TB was significantly more common [182]. Similarly, it has been found in an other study that (63.5%) of the MDR-TB cases were from 15 to 44 years of age and was marginally statistically associated with MDR-TB [183]. Several studies showed the absence of statistically significant difference in the proportion of any resistance by age [184]. Moreover gender does not appear to make any significant difference on drug resistance TB. Numerous studies have reported that females had a higher risk than males for MDR-TB [185, 186]. Analysis in one of the study showed that there was no evidence of an association between sex and risk of MDR/RR-TB in TB patients both globally and nationally in the majority (81%, 86 out of 106) of countries, with an overall random-effects weighted M:F risk ratio of 1.04 (95% CI 0.97–1.11) [187]. Supplementary files in Appendix D provide additional graphs for reference. These findings contribute to understanding the demographic patterns and outcomes associated with different types of drug-resistant TB.

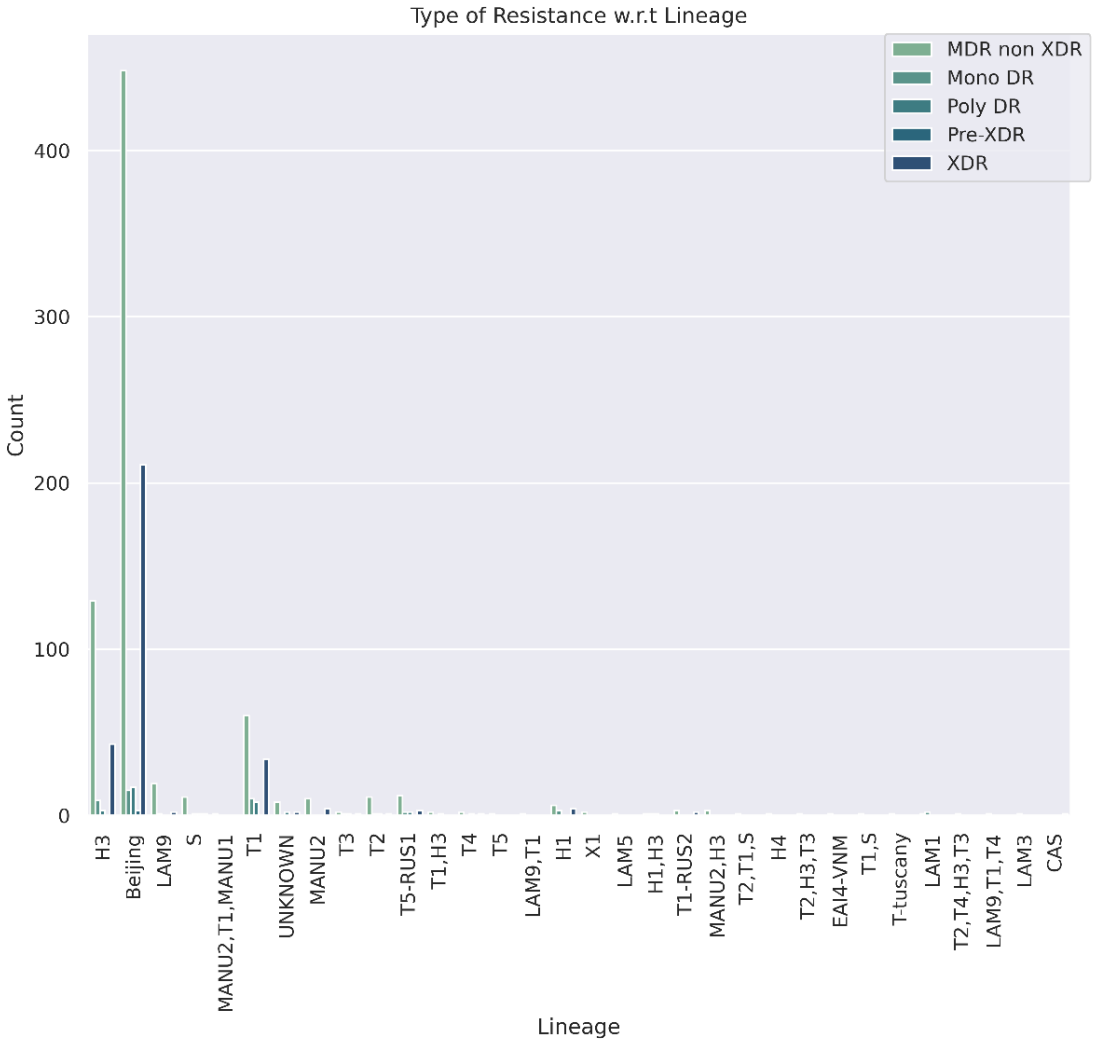


FIGURE 4.7: Bar chart demonstrating frequency of each type of resistance against the lineage.

4.2 Pattern Identification of Multidrug Resistant Isolates with Respect to Drug Susceptibility Through Data Mining

Drug susceptibility test results dataset from TB portal consisted of 1970 records and more then 4500 rules were generated by implementing aproiri algorithm in python script. The optimum combination for threshold values of minimum support and confidence was 0.01 and 1. The Most rules had high confidence ,indicating that the antecedent nearly perfectly predicted the consequent. The confidence threshold for pruning the rules was set at 0.9 to include rules with a reliability

of 90%. The support of most rules was small ,which is consistent with the low frequency of resistance to most antimicrobials other than isoniazid and rifampin. Pruning with confidence 0.9, support >0.01, results around 500 best-rules in the dataset. Lift compares the support of a rule in a dataset to the support expected if the antecedent and consequent were independent.

A lift of 1 indicates that the antecedent and consequent are independent; a lift 1 is a positive association. 70% of rules across all datasets had a negative association between the antecedent and consequent means 70% of resistance were dependent and 30% was dependent. The generated ruleset was visualize and filtered manually as it contained a lot of vague and meaningless rules. A rule is considered meaningless with respect to the study if it contain more than 3 items in antecedent.For example the rule shown in Table 4.1 is insignificant as it depicts the in combination [KAN, INH, CAP, OFL, STR, PTH, RIF, EMB] exist when the combination [RIF, OFL, LVX, AMK] will show up. This does not make any sense as most of drugs are overlapping in both rules. Another fact that support of such rules were really low hence a number of such rules did not meet the criteria. Moreover, the rules with zero antecedent and consequent were eliminated. Anyhow some of the itemset are were important with respect to the nature of the data. Meaningful rules were filtered and classified according to the drugs and types of resistance. The selected rules were with highest support, with antecedent having not more items than 3,with no or minimal overlap of items in antecedent and consequent. The most frequent rule was with the maximum support of 0.5 was [RIF to INH].This reflects the high degree of coexistence rifampicin [RIF] and isoniazid [INH] resistance in the dataset.(Table 4.2).This rule complemented the definition of multidrug resistance. The target genes responsible for INH resistance *katG* and *inhA* are significantly related to mutations in *rpoB* that is responsible for RIF resistance. The second most credible rules were those representing extensive drug resistance i.e the MDR paired with one or more of the second line drugs (SLD) resistance. These rules had RIF,INH in antecedent and two or more SLD in consequents with the support ranging from 0.01 to 0.08. The association rules with

TABLE 4.1: Eliminated association rule due to low support value.

Antecedents	**Consequent**	**Support**	**Conf.**
KAN, INH, CAP, OFL, STR, PTH, AMK RIF, EMB	RIF, OFL, LVX,	0.011	1

TABLE 4.2: Eliminated association rule due to low support value.

Antecedents	**Consequent**	**Support**	**Lift**
RIF	INH	0.5	1

two and three drugs in consequents are enlisted in table below. Rest of the rules are given in Appendix E.

The rules with support equal or greater than 0.05 contained the second line drugs Capreomycin (CAP), Ofloxacin (OFL), Ethambutol (EMB) Amikacin (AMK) and Kanamycin (KAN) in the consequent. The targets genes against these drugs are *g.TlyA, g.gyrA, g.gyrB, g.embA, g.embB, g.embC* and *g.rrs* are closely related in terms of mutation and developing resistance.

In conclusion the EDA was followed by datamining technique on the DST data of the same patients. The association rule mining was applied to drive association in the context of drug resistance in TB treatment. An extensive set of over 4500 association rules was generated, with the objective of understanding drug resistance patterns, particularly within the context of TB treatment with minimum support of 0.01 and confidence of 0.9. The rules were filtered eliminating vague and irrelevant rules, focusing only on those of significance. These significant rules were then classified based on the drugs involved and the types of resistance they represented. The most frequent and impactful rule, with boasting the highest support value of 0.5 was representing the correlation between resistance to isoniazid (INH) and rifampicin (RIF). Two basic first-line drugs in the against TB. This

TABLE 4.3: Top 20 association rules complementing XDR.

Antecedents	**Consequent**	**Support**
RIF, INH	CAP, OFL, STR	0.076142132
RIF, INH	CAP, OFL, STR	0.076142132
RIF, INH	CAP, EMB, OFL, STR	0.070050761
RIF, INH	OFL, AMK, STR	0.059898477
RIF, INH	OFL, EMB, AMK, STR	0.055837563
RIF, INH	OFL, EMB, AMK, STR	0.055837563
RIF, INH	KAN, CAP, OFL, STR	0.052791878
RIF, INH	STR, CAP, AMK, OFL	0.052284264
RIF, INH	KAN, CAP, OFL, STR, EMB	0.050761421
RIF, INH	AMK, CAP, OFL, STR, EMB	0.049238579
RIF, INH	KAN, OFL, AMK, STR	0.041624365
RIF, INH	KAN, AMK, OFL, STR, EMB	0.040609137
RIF, INH	KAN, AMK, CAP, OFL, STR	0.039593909
RIF, INH	KAN, AMK, CAP, OFL, STR, EMB	0.039086294
RIF, INH	CAP, EMB, PTH, STR	0.025888325
RIF, INH	CS, EMB	0.022335025
RIF, INH	EMB, PTH, AMK, STR	0.022335025
RIF, INH	CAP, LVX, STR	0.021827411
RIF, INH	KAN, PZA, EMB	0.021319797
RIF, INH	KAN, PTH, STR	0.021319797
RIF, INH	CAP, PTH, OFL, STR	0.020812183
RIF, INH	LVX, AMK, STR	0.020304569

rule essentially serves as phenotypic evidence, providing empirical proof of the link between INH and RIF resistance. The other frequent associations TB drugs, particularly rifampicin (RIF) and isoniazid (INH) in the antecedents, and multiple second-line drugs (SLD) (Capreomycin (CAP), Ofloxacin (OFL), Ethambutol (EMB), Amikacin (AMK), and Kanamycin (KAN) in the consequents. These drugs are essential in treating drug-resistant TB strains.The target genes as per literature associated with these second-line drugs, including *TlyA, gyrA, gyrB, embA, embB, embC* and *rrs* were expected to be somehow related to each other . Furthermore, the outcomes of data mining laid the foundation for next genotypic analysis, which involved identifying resistance genes within the genomic sequences of patients, previously identified as having phenotypic resistance through data mining and exploratory data analysis (EDA). The association rules supported the statistical trends observed in EDA, with the highest proportion of cases being multidrug-resistant (MDR), followed by extensively drug-resistant (XDR) cases and other types of resistances. Additionally, association rules generated from the data also provided insights into the specific drug combinations associated with MDR and XDR cases. These resistant drug combinations basically corresponded to their targeted genes developed resistance to the drugs due to some mutation mechanisms.

4.3 Identification of Novel AMR Genes in MDR and XDR TB

Utilizing XDR and MDR isolates samples, an investigation was conducted to identify the genetic elements responsible for antibiotic resistance using the whole genome sequencing pipeline, Ariba. The output was csv files results that include information such as gene names, variant types, novelty of the variants, and the effects of the variants on the gene sequence. The purpose of analysis was exploring the presence of novel coding gene variants that could potentially be linked

to antimicrobial resistance in two types of *M. tuberculosis* strains, namely extensively XDR and MDR strains. This analysis unveiled previously known and newly discovered variants of these resistance genes. The novel variants were specifically selected for further scrutiny .A comparative analysis performed to identify the shared novel variants in XDR and MDR novel resistance genes, revealing a set of unique novel variants exclusive to each category. Notably, one unique novel variant was exclusively found in the XDR samples, while 16 were exclusively identified in the MDR and XDR .Moreover, 29 variants were common between the XDR and MDR samples as given. After literature verification it was revealed that 12 of the genes were not from *M. tuberculosis*. Such genes were place in another cetagory. After this filteration 6 Unique novel variants genes in MDR TB were left and 27 common XDR TB/MDR TB novel variants' genes while 12 were Novel unidentified variants gene not reported for TB drug resistance As given in Table 4.6.

4.3.1 Unique Novel Variants Conferring MDR Resistance

The results from ARIBA revealed the presence of six frequently occurring novel variants in the data, namely *g.msrA, g.inhA, g.rpoC, g.mfpA, g.fusA,* and *g.qac*, which confer either MDR exclusively. Other genes have also been reported in literature but these six were found on the Genomic sequences from the TB Portals. Table 4.4 contains the genes functions and type of resistance from each gene.

g.msrA helps repair proteins that have been damaged by oxidative stress, including those targeted by reactive oxygen species such as hydrogen peroxide. By restoring the functionality of these proteins, MsrA contributes to the bacterium's ability to adapt to and survive in the presence of oxidative stress [194].

Involvement of msrA is in drug resistance is not frequently reported in literature anyhow It was reported that *msrA* was one of the genes that reduced INH susceptibility.A mutation in *g.msrA*, could impact the *g.inhA* binding [195].

TABLE 4.4: Identified variants' genes and their functions

Sr. No.	Gene	Protein Name	Function	Ref.
1	*g.msrA*	Peptide methionine sulfoxide reductase	L-methionine: thioredoxin-disulfide-S oxidoreductase activity	[188, 189]
2	*g.inhA*	Enoyl-[acyl-carrier-protein] reductase	Synthesis of fatty acids, specifically mycolic acid	[190]
3	*g.rpoC*	DNA directed RNA polymerase subunit Beta	catalysis of the transcription process using ribonucleotide phosphates as substrates	[191]
4	*g.mfpA*	Pentapeptide repeat protein	Inhibition of ATP-independent relaxation of DNA	
5	*g.fusA*	Elongation factor G	Catalysis of GTP-dependent ribosomal translocation during the process of translation	[192]
6	*g.qac*	Qac A/B Quaternary Ammonium Compound Efflux MFS transporter	Transmembrane transporter activity	[193]

Resistance mechanisms associated with the inhA gene play a significant role in multidrug-resistant TB typically involve genetic mutations that undermine the efficacy of INH.One primary mechanism involves changes in the INH binding site on the InhA enzyme, encoded by the *g.inhA* gene. These mutations result in reduced affinity between INH and InhA, diminishing the drug's ability to inhibit the synthesis of mycolic acids, vital components of the mycobacterial cell wall. Some INH-resistant strains of *M. tuberculosis* exhibit mutations in the promoter region of the inhA gene.

Notably, the substitution at position -15 in the inhA promoter can lead to the over-expression of the InhA enzyme. Although this mutation doesn't directly modify the drug's binding site, increased *inhA* enzyme levels compensate for the reduced affinity of INH, rendering the drug less effective [44, 196]. Mutations in the *rpoC* gene are important in making TB resistant to multiple drugs. TB can become resistant to a drug called rifampin because of changes in the *rpoB* gene. But these changes can affect how well the bacteria can survive. To balance this, the bacteria can also change other genes like *rpoA* or *rpoC*. A study in China looked at these gene changes in TB samples from patients.

They found that in a certain type of TB, around 28% of the samples that were resistant to rifampin had changes in the rpoA or *rpoC* gene. These changes were more common in new cases and when combined with another change. The TB strains with changes in rpoC were also connected to specific patterns. These findings show that these gene changes play a big role in making TB resistant to drugs and in how it spreads [197]. *mfpA* is associated with the bacterium's cell wall structure and may play a role in biofilm formation. The modification in the cell wall structure and components like *mfpA* can possibly alter drug susceptibility.

Anyhow the main mechanisms of drug resistance in *M. tuberculosis* imply mutations in genes directly related to drug targets and activation. *mfpA* interacts with the DNA gyrase enzyme, crucial for DNA replication. This interaction prevents fluoroquinolone antibiotics from binding to the enzyme, potentially contributing

to drug resistance. The *mfpA* gene is present in both drug-sensitive and drug-resistant *M. tuberculosis* strains, with a higher frequency in drug-resistant strains, as it can be acquired through horizontal gene transfer. *fusA* is involved in protein synthesis. It is associated with elongation factor G (EF-G), which facilitates the translocation of ribosomes during translation.

Mutations in the *fusA* gene can lead to resistance to fusidic acid, an antibiotic that inhibits bacterial protein synthesis by binding to elongation factor G (EF-G) [192].

Multiple invitro testing proved that fusidic acid is potential anti TB drug that can be included in first line treatment regime for TB treatment [198]. Lim *et al.*performed a study in 2014 to show that rpoB and fusA causes resistance against rifampcin and fusidic acid in bacteria [199, 200].

4.3.2 Common Novel Variants Conferring MDR/XDR Resistance

The List with samll l comprises 27 frequently occurring common variants in the data, including *g.thyA, g.embR, g.embB, g.embC, g.folC, g.kasA, g.aac, g.mshA, g.iniA, g.rpsA, g.gyrA, g.gidB, g.iniC, g.pncA, g.ethA, g.Rv1258c, g.embA, g.murA, g.iniB, g.ribD, g.gyrB, tlyA, efpA, rpsL, g.katG, g.rpoB,* and *g.blaC.*

These genes are associated with both MDR and XDR TB. Table 4.5 provides details of the gene functions and the type of resistance they confer.

A literature survey was conducted to verify the involvement of each gene in antimicrobial drug resistance TB.

Mutations in the *g.embA, g.embB, g.embC, g.embR, g.katG, g.pncA, g.thyA* and *g.rpsL* genes that are associated with phenotypic resistance in various anti-TB drugs [38].

TABLE 4.5: Common novel variants' genes and their functions

Sr. No.	Gene Name	Protein Name	Function	Ref.
1	*g.thyA*	Thymidylate synthase ThyA	Thymidylate synthase activity	[45]
2	*g.embR*	Transcriptional regulatory protein	Positive regulation of embCAB operon transcription Have GTPase and ATPase activity	[201]
3	*g.embB*	Probable arabinosyltransferase B	Acts to polymerize arabinose into arabinan	[202]
4	*g.embC*	Probable arabinosyltransferase C	Involves in the polymerization of arabinose into arabian by forming alpha(1-5) linkage	[203]
5	*g.folC*	Dihydrofolate synthase	Dihydrofolate synthase activity	[204]
6	*g.kasA*	3-oxoacyl-[acyl-carrier-protein] synthase 1	Involves in the mycolic acid biosynthesis process	[205]
7	*g.aac*	Aminoglycoside 2'-N-acetyltransferase	Catalysis of coenzyme A dependent acetylation of 2' hydroxyl of aminoglycosides	[206]
8	*g.mshA*	D-inositol 3-phosphate glycosyltransferase	Catalyzes mycothiol biosynthesis	[207]
9	*g.iniA*	Isoniazid-induced protein iniA	Efflux pump for drug tolerance	[208]
10	*g.rpsA*	30S ribosomal protein S1	Binding with mRNA and facilitates its recognition by 30S ribosomal subunit during translation initiation step	[209]

Sr. No.	Gene Name	Protein Name	Function	Ref.
11	*g.gyrA*	DNA gyrase subunit A	Type II topoisomerase that negatively supercoils DNA which closed and circular	[210]
12	*g.gidB*	Ribosomal RNA small subunit methyltransferase G	A methyltransferase enzyme that adds methyl group in 16S ribosomal RNA	[211]
13	*g.iniC*	Isoniazid Inducible Protein	Peroxidase activity	[212]
14	*g.pncA*	Nicotinamidease/ pyrazinamidase	Involves the deamination of nicotinamide into nicotinate	[213]
15	*g.ethA*	FAD containing monooxygenase	Conversion of a wide range of ketones into lactones or esters via Baeyer-Villiger reaction	[214]
16	*g.Rv1258c*	Multidrug Efflux Pump	Acts as an efflux pump for a variety of proteins and contribute to antimicrobial drug resistance	[215]
17	*g.embA*	Probable arabinosyltransferase A	Acts to polymerize arabinose into arabinan	[202]
18	*g.murA*	UDP-N-acetylglucosamine 1-carboxyvinyl transferase	Responsible forcell wall formation by adding enolpyruvyl to UDP-N-acetylglucosamine	[216]
19	*g.iniB*	Isoniazid induced protein	Drug resistance by efflux	[217]
20	*g.ribD*	Riboflavin biosynthesis protein	Diaminohydroxy phosphoribosylamino pyrimidine deaminase activity	[218]

Sr. No.	Gene Name	Protein Name	Function	Ref.
21	*g.gyrB*	DNA gyrase subunit B	Assist in supercoiling of relaxed DNA in an ATP-1independent manner	[219]
22	*g.tlyA*	16S/23S rRNA (cytidine-2'-O) methyltransferase	Functions as a host evasion factor leading to pathogenesis	[220]
23	*g.efpA*	Integral membrane efflux protein	Involves in the transmembrane transporter activity	[221]
24	*g.rpsL*	30S ribosomal protein S12	Interacts with S4 and S5 and regulates translational activity	[222]
25	*g.katG*	Catalase-peroxidase	A bifunctional protein with catalase and peroxidase activity	[223]
26	*g.rpoB*	DNA directed RNA polymerase subunit beta	Catalyzes transcription of DNA into RNA by utilizing ribonucleotide triphosphates	[224]
27	*g.blaC*	Beta-lactamse	Inactivates beta lactam antimicrobial drugs by hydrolyzing the amide group	[225]

A study using transposon mutagenesis identified mutations in the thyA gene associated with resistance to PAS that were also present in clinical isolates resistant to PAS. thyA gene gene encodes thymidylate synthase, a which act as a key enzyme in DNA synthesis.Mutations in thyA can lead to reduced enzyme activity, altering the balance of nucleotides necessary for DNA replication [226]. Mutations in the embR gene play a role in resistance to Ethambutol (EMB) in drug resistant TB. as encodes a transcriptional regulator thatregulates the expression of the embCAB operon, which is involved in the biosynthesis of arabinogalactan, an essential component of the mycobacterial cell wall. Arabinogalactan is crucial for the structural integrity of the cell wall, and disrupting its synthesis can lead to cell wall abnormalities and drug resistance.Mutations in *embR* can lead to upregulate expression of the embCAB operon, which can result in changes to the cell wall structure.This affacts the permeablility cell wall to EMB or altering the target sites of the drug [227].

The *embA, embB* and *embC* genes collectively form the embCAB operon in Mycobacterium TB which is integral to the biosynthesis of the mycobacterial cell wall. Each of these genes plays a distinct yet interconnected role in this crucial process. *embA* and *embC* are involved in the synthesis of arabinogalactan, a vital component of the mycobacterial cell wall, while embB plays a central role in the incorporation of arabinogalactan into the cell wall structure. Mutations in these genes can lead to resistance to ethambutol (EMB), an essential anti-TB drug. Specifically, mutations in *embB* are most associated with EMB resistance, affecting the drug's binding site and inhibitory action on cell wall synthesis. *embA* and *embC* mutations can also contribute to EMB resistance by disrupting arabinogalactan synthesis. Collectively, these genes are pivotal in maintaining the structural integrity of the mycobacterial cell wall and are key players in the development of drug resistance in TB. Understanding their functions and mutations is critical for diagnosing and addressing drug-resistant TB effectively [227].

In para-aminosalicylic acid (PAS)-resistant strains of *M. tuberculosis*, specific mutations in the *folC* gene have been reported , such as I43T, I43A, and E40G.These

mutations are responsible for resistance to PAS, as they increases MIC (Mininmum inhibitory concentrations) to inhibit bacterial growth. Structural studies of FolC reveal that these mutations alter the substrate-binding pocket of the enzyme. exhibit reduced enzymatic activity. When the wild-type folC gene is reintroduced into these PAS-resistant strains, their susceptibility to PAS is restored, confirming the role of folC mutations in PAS resistance [16].

KasA enzyme involved in mycolic acid synthesis.Mycolic acids contribute to the impermeability and structural integrity of the bactaium cell wall. KasA shares some similarities with InhA. Mutations in kasA consequently, changes in the mycobacterial cell wall structure and permeability. These alterations could affect drug permeability and susceptibility.

It has been proposed that kasA might serve as a secondary target of INH, with the primary target being the enzyme InhA [228].

Aminoglycoside 2'-N-acetyltransferase (AAC) enzymes are involved in modification of aminoglycoside antibiotics, such as kanamycin and amikacin. Bacterial strains of *M. tuberculosis* that produce AAC enzymes can acetylate these antibiotics, making them ineffective. This result in the structural changes in the drug, reducing its ability to bind to the ribosomal target in the bacterial cell, where it normally interferes with protein synthesis. As a result, the modified antibiotic is less effective at inhibiting bacterial growth [229].

The mshA gene in encodes an enzyme involved in mycothiol synthesis. In *M. tuberculosis*, Mycothiol is involved in tolerance against oxidative stress and toxic. Mutations in mshA have been associated with drug resistance in TB, particularly isoniazid (INH) resistance. These mutations can disrupt mycothiol synthesis, affecting the bacterium's ability to neutralize reactive oxygen species generated by INH. MshA mutations contribute to INH resistance and may play a role in drug tolerance and multidrug resistance in *M. tuberculosis* [230].

In Mycobacterium TB, Isoniazid inductible gene protein iniA may play a role in a naturally occurring tolerant phenotype in clinical *M. tuberculosis* infections, potentially affecting the duration of TB treatment. IniA, IniB, and IniC are integral components of the iniBAC operon. IniA functions as a pump component and may help *M. tuberculosis* survive antibiotic exposure by preserving cellular functions or removing toxic components of the cell wall. Colangeli et al. [208], IniB also participates in the bacterial response to antibiotic stress and contributing to antibiotic resistance. IniC, another component of the operon, similarly plays a role in drug resistance by responding to antibiotic stress, including INH exposure,. Collectively, these three genes are crucial in understanding the mechanisms underlying antibiotic resistance in *M. tuberculosis*, particularly against INH, aiding in the development of diagnostic and treatment strategies for TB [231].

Ribosomal protein S1 (RpsA) plays a role resistance to the antibiotic pyrazinamide (PZA) in TB. Mutations in the *rpsA* gene, can lead to modifications in the enzyme pyrazinamidase (PZase) activity. PZA requires activation into pyrazinoic acid (POA) to be effective against *M. tuberculosis*. *rpsA* mutations can affect PZase activity, reducing PZA activation, and resulting in PZA resistance. However, not all PZA-resistant *M. tuberculosis* strains have rpsA mutations, as resistance can also involve mutations in the pncA gene, which directly affects PZase production. Mutations in *pncA* can reduce or eliminate PZase activity, rendering PZA ineffective and contributing to drug-resistant TB [232].

Mutations in the *gyrA* and *gyrB* genes are associated with resistance to fluoroquinolone drugs in TB . *gyrA* mutations can prevent the fluoroquinolone from binding effectively to the DNA gyrase,. Mutations in *gyrB* can lead to structural changes in the DNA gyrase complex, making it less susceptible to fluoroquinolones. DNA gyrase is essential for DNA replication and repair in *M. tuberculosis*, the bacterium causing TB. Mutations in *gyrA* and *gyrB* can result in cross-resistance to multiple fluoroquinolone drugs, limiting treatment options for drug-resistant TB patients. drug- to fluoroquinolone antibiotics. Detecting these mutations is crucial for diagnosing drug resistance and guiding treatment decisions in TB patients [233].

The *gidB* gene encodes a methyltransferase enzyme that plays a crucial role in bacteria by methylating a specific nucleotide position within the 16S ribosomal RNA (rRNA). This modification helps regulate the structure and function of the ribosome, which is essential for protein synthesis. GidB's function is to fine-tune ribosomal activity, ensuring efficient protein production and adaptation to different environmental conditions in bacteria. Streptomycin is an antibiotic that binds to the ribosome during protein synthesis in bacteria. It interacts with the 16S rRNA of the ribosomal subunit, specifically at a site where gidB-mediated methylation typically occurs [234].

ethA and *ethR* are key components of drug-resistant TB. *ethA*, encoded by the enzyme is responsible for activating the antibiotic ethionamide(ETH). *ethA* converts ethionamide into an active form that disrupts the TB bacterium's cell wall synthesis, ultimately leading to bacterial cell death.whereas *ethR* acts as a transcriptional repressor that regulates the expression of *ethA*. When EthR binds to the ethA promoter region, it inhibits *ethA* transcription, reducing ethionamide activation. Mutations or alterations in these genes and their associated proteins can influence the efficacy of ethionamide and contribute to drug resistance in TB. Understanding the roles of *ethA* and *ethR* is essential for planning strategies to combat drug-resistant TB effectively [235].

Rv1258c is associated with efflux-mediated drug resistance in TB, particularly rifampicin resistance. Efflux pumps are cellular mechanisms that pump drugs and antibiotics out of bacterial cells, reducing their effectiveness. It is involved in pumping rifampicin out of bacterial cells, reducing the drug's effectiveness . Understanding the role of Rv1258c is important for developing strategies to combat drug-resistant TB.

Mur enzymes (MurA-MurF) are involved in the peptidoglycan synthesis pathway of *M. tuberculosis* which are building blocks the cell wall of *M. tuberculosis*. inhibition Mur enzymes may lead to disruption of the cell wall synthesis of *M. tuberculosis*, making the it more vulnerable to antibiotics.So Mur enzymes potential

as targets for drug development as these enzymes could be a promising strategy for developing new drugs to combat drug-resistant TB. Anyhow the resistant mechanism of these enzymes are not reported yet [236].

In *M. tuberculosis* ,*ribD* is involved in the folate metabolism pathway. Mutations in *ribD* can cause resistance to the anti-TB drug para-aminosalicylic acid (PAS). Some mutations, can lead to overexpression of *ribD* and increased production of tetrahydrofolate (THF), which is a critical cofactor in various metabolic processes. This overproduction of THF can reduce the inhibitory effects of PAS on folate metabolism, contributing to drug resistance.

The *tlyA* gene in *M. tuberculosis* encodes a methyltransferase enzyme that is involved methylation of ribosomal RNA (rRNA), particularly the 16S rRNA results in functional regulation of ribosome.These alterations in tlyA-mediated methylation can affect antibiotic resistance and the overall fitness of *M. tuberculosis* strains. Methylation by *tlyA* can lead to resistance to certain antibiotics, such as aminoglycosides, and may compensate for fitness costs associated with resistance mutations [237].

efpA is a gene found in *M. tuberculosis* (the bacterium that causes TB) that When exposed to drugs like isoniazid (INH) and other compounds, the expression of *efpA* increases, indicating its potential role in drug resistance. contributes to the bacterium's ability to respond to antibiotic treatment [238]. The *rpsL* gene encodes ribosomal protein S_{12}, a vital component of the bacterial ribosome involved in the translation of mRNA . Hence The *rpsL* gene in Mycobacterium TB plays a vital role in protein synthesis. Mutations in rpsL are a common mechanism of resistance to the antibiotic streptomycin in TB. These mutations, reduce streptomycin's binding affinity to ribosomal protein S_{12} [44].

katG in *M. tuberculosis* plays a pivotal role in activating isoniazid (INH), a critical antibiotic used in TB treatment. This enzyme catalyzes the conversion of INH into its active form, disrupting mycobacterial cell wall synthesis. Resistance to INH often arises from mutations in the *katG* gene, hindering KatG's ability to activate INH effectively, reducing the drug's potency against TB bacteria [237].

The *rpoB* gene in Mycobacterium TB encodes the beta subunit of RNA polymerase, a key enzyme involved in bacterial transcription. It plays an important role in initiating transcription, synthesizing RNA from DNA, and regulating gene expression. Mutations in *rpoB* are a usual reason of resistance to the antibiotic rifampicin (RIF) . These mutations interfere with RIF binding, leading to drug resistance [239].

The *blaC* gene in *M. tuberculosis* encodes a β-lactamase enzyme, which is a major mechanism of resistance to β-lactam antibiotics like penicillin [240].

4.3.3 Novel Unidentified Variants Not Reported For TB Resistance

The ARIBA analysis for genomic data revealed 12 genes that have not been reported in literature with respect to *M. tuberculosis*. However they have been associated with resistance mechanisms in other bacterial species. The presence of these sequences in TB isolates from TB portals suggests their potential involvement in resistance mechanisms in *M. tuberculosis*. For this reason these genes are also subjected to functional enrichment analysis.

The shortlisted XDR and MDR strains were then analyzed through ARIBA pipeline to identify novel variants in each isolate. The study's significant findings shed light on the variants associated with MDR and XDR strains, potentially playing a crucial role in TB drug resistance. The investigation successfully identified variants specifically linked to MDR and XDR, suggesting their direct or indirect involvement in the development of drug resistance.

A total of 27 common genes with variants were observed in MDR and XDR samples including *g.thyA, g.embR, g.embB, g.embC, g.folC, g.kasA, g.aac, g.mshA, g.iniA, g.rpsA, g.gyrA, g.gidB, g.iniC, g.pncA, g.ethA, g.Rv1258c, g.embA, g.murA, g.iniB, g.ribD, g.gyrB* and *g.tlyA*. However the 6 MDR genes identified including*g.msrA, g.inhA, g.rpoC, g.mfpA, g.fusA, g.qac*. Certain genes which are associated with XDR TB overlap with MDR ,as XDR-TB is characterized by resistance to both

TABLE 4.6: Novel unidentified variants not reported for TB resistance.

Sr. No.	Gene	Protein Name	MDR/XDR	Function	Ref.
1	*g.parC*	DNA topoisomerase 4, subunit A	MDR	Chromosmal Segregation, catenation of newly synthesized chromosomes	[241, 242]
2	*g.blaZ*	Beta-Lactamase	MDR	Act as a reporter for protein exports during the mycobacterium TB infection	[243]
3	*g.mecA*	Adapter protein MecA 1	MDR	Responsible for chaperone activity by recognizing unfolded and aggregated proteins	[244]
4	*g.mecI*	Methiciline resistance regulatory protein	MDR	Represses the transcription of gene encoding for penicillin binding protein *mecA*	[245]
5	*g.tetA*	Tetracycline resistance protein A	MDR	An efflux protein that allows resistance against tetracycline by decreasing its accumulation in cell wall	[246]
6	*g.mfpA*	Pentapeptide repeat protein	MDR	Inhibition of ATP-independent relaxation of DNA	
7	*g.inuA*	Inulinase	MDR	Hydrolyzing enzyme which breaks down inulin into fructose by acting on beta 2, 1 linkages	[247]
8	*g.aadA27*	Aminoglycoside (3") (9) adenylyl transferase	MDR	Involves the adenylyl transferase activity	[248]
9	*g.ANT*	Non-toxic nonhemag lutinin type A	MDR	Functions with botulinum neurotoxin type 1 to protect it against inactivation due to pH	[249]

Sr. No.	Gene	Protein Name	MDR/XDR	Function	Ref.
10	*g.TEM*	Beta Lactamase	XDR	Beta lacatamse activity	[250]
11	*g.tuf*	Elongation factor Tu	MDR and XDR	GTPase activity and GTP binding	[251]
12	*g.erm*	rRNA adenine N-6-methyl transferase	MDR and XDR	Acts to produce dimethylation at adenine residue at position 2085 of 23S *rRNA* and minimize the affinity between ribosomes and antibiotics	[252]

first-line and second-line anti-TB drugs, making it more complex. These genes are associated with resistance to different drugs like *g.katG, g.inhA, g.aphC, g.kasA, g.rpsL, rrs, g.embB, g.pncA, g.gyrA, g.gyrB, g.eis, g.tlyA, g.whiB7,* efflux pump genes. Various genes related to efflux pumps can contribute to drug resistance in TB, various other mutations and genes associated with resistance to second-line drugs like capreomycin, moxifloxacin, and others.

It has been reported from multiple other studies that individuals who have previously undergone TB treatment are more likely to develop MDR-TB compared to new TB cases. These findings emphasize the importance of ensuring treatment completion rates for new cases and maintaining a vigilant approach to monitoring drug resistance in individuals with a history of previous TB treatment. Furthermore, these insights underscore the critical need for rigorous implementation of infection control measures to prevent the transmission and emergence of new cases, including those with drug-resistant TB [253, 254].

The incidence of XDR-TB remained low and showed no significant change during the study period, there was a notable increase in the incidence of pre-XDR TB. This rise in second-line drug resistance is a matter of concern. It has the potential to lead to important implications for case management, including the adjustment of treatment regimens, the demand for new therapeutic agents, and the introduction of rapid diagnostic tools. Previous studies worldwide have also reported a similar upward trend in second-line drug resistance [255].

Overall, this analysis provides valuable insights into the genetic factors contributing to drug resistance in TB, helping researchers and healthcare professionals better understand and manage drug-resistant TB strains. It also sets the stage for further genetic validation, where the identified associations will be scrutinized at the genotypic level, offering deeper insights into the underlying genetic mechanisms driving drug resistance in TB.

In order to understand the significance of association rules in previous step the list of identified genes were compared with the specified targets of drugs present in top rules and most frequent itemset from the association rules mining.For instance

the coexistence of mutations in the *rpoB, KatG* and *inhA* genes, as listed in the Table 3.1, is characteristic of MDR TB combinations. In accordance with the first rule from table 4.3 [RIF, INH]=>[CAP, OFL, STR], these drugs target the combination of resistance genes *rpoB, KatG*, and *inhA*, which subsequently leads to resistance to *g.tlyA, g.gyrA/g.gyrB, g.rrs*, and *g.rpsl*. This combination of resistance is associated with MDR/XDR TB, as it involves resistance to first-line drugs (RIF and INH) as well as second-line drugs (STR, CAP, and OFL) all four of these gene existed list of identified either common or unique variant genes from ARIBA pipeline. Similar combinations found throughout the table 4.3 correspond to various XDR mechanisms and the coexistence of these mutations. In this way both results showed more than 70% similarity.

4.4 Functional Enrichment Analysis

Functional enrichment analysis was conducted using the gene symbols of these novel variants in order to identify the pathways and functions that exhibited enrichment. Two distinct functional enrichment analyses were carried out. The first analysis focused on the common genes shared between MDR and XDR, while the second analysis specifically targeted the unique genes associated with MDR. The former analysis involved 31 common novel variant genes found in both MDR and XDR, whereas the latter analysis considered the genes associated with novel unique variants that distinguished MDR from XDR.

The functional enrichment analysis of common novel variant genes revealed that 15 genes were significantly associated with the antimicrobial resistance pathways. These genes, namely *g.gyrB, g.gyrA, g.aac, g.iniA, g.tap, g.rpoB, g.rpsL, g.katG, g.pncA, g.blaC, g.floC, g.thyA, g.embC, g.embA*, and *g.embB*, exhibited a strong enrichment signal (strength = 1.6) and a very low false discovery rate (FDR = 1.53e-17). Through STRING network analysis it was elucidated that these genes form a network of biological interactions as shown in Figure 4.8.

Furthermore, several other functional terms were identified for the remaining 16 common genes through the enrichment analysis. For example, seven genes involved in RNA-binding were found, including *g.rpsF, g.rplI, g.rplK, g.rplA, g.rpsL, g.rpsA*, and *g.tlyA*. Additionally, six genes associated with ribosomal proteins were detected, namely *rpsF, rplI, rplK, rplA, rpsL*, and *rpsA*. The analysis also revealed five genes related to rRNA-binding (*g.rpsF, g.rplI, g.rplK, g.rplA, and g.rpsL*), five genes associated with cell wall biogenesis/degradation (*g.murB, g.murA, g.embC, g.embA, and g.embB*), two genes involved in topoisomerase activity (*g.gyrB* and *g.gyrA*), four genes associated with glycosyltransferase activity (*g.mshA, g.embC, g.embA*, and *g.embB*), and eleven genes related to transferase activity (*g.aac, g.mshA, g.rpoB, g.murA, g.tlyA, g.kasA, g.thyA, g.embC, g.embA, g.embB*, and *g.gid*).

The Gene Ontology analysis revealed, 11 genes were observed to be associated with GO:0003676 (Nucleic acid binding), including *g.gyrB, g.gyrA, g.rpsF, g.rplI, g.rplK, g.rplA, g.rpoB, g.rpsL, g.embR, g.rpsA*, and *g.tlyA*. The strength of the enrichment signal was 0.61, with a false discovery rate (FDR) of 0.0069. GO:0003723 (RNA binding): 7 genes, namely *g.rpsF, g.rplI, g.rplK, g.rplA, g.rpsL, g.rpsA*, and *g.tlyA*, were identified as being associated with RNA binding. The enrichment had a strength of 0.89 and an FDR of 0.0069. GO:0003735 (Structural constituent of ribosome): 6 genes (*g.rpsF, g.rplI, g.rplK, g.rplA, g.rpsL*, and *g.rpsA*) were found to be associated with this term. The strength of the enrichment was 1.11, with an FDR of 0.0069. GO:0019843 (rRNA binding): This term was associated with 5 genes, including *g.rpsF, g.rplI, g.rplK, g.rplA*, and *g.rpsL*. The strength of the enrichment was 1.17, with an FDR of 0.0069. Additionally, several other terms showed significant associations with the common novel variant genes. These included GO:0052636 (Arabinosyltransferase activity), GO:0097159 (Organic cyclic compound binding), GO:1901363 (Heterocyclic compound binding), GO:0034335 (DNA negative supercoiling activity), GO:0071949 (FAD binding), GO:0005488 (Binding), GO:0016740 (Transferase activity). Based on Gene Ontology Biological Processes enrichment results for the common genes, it was found that 14 genes were associated with the term "Response to antibiotic" (GO:0046677).

These genes include *g.gyrB, g.gyrA, g.aac, g.iniA, g.rpoB, g.rpsL, g.tap, g.pncA, g.blaC, g.folC, g.thyA, g.embC, g.embA*, and *g.embB*. The strength of this enrichment signal was 1.45, with a false discovery rate (FDR) of 1.61e-13. Furthermore, 15 genes were identified to be associated with the term "Response to chemical" (GO:0042221), including *g.gyrB, g.gyrA, g.aac, g.iniA, g.rpoB, g.rpsL, g.tap, g.katG, g.pncA, g.blaC, g.folC, g.thyA, g.embC, g.embA*, and *g.embB*. The strength of this enrichment was 1.04, with an FDR of 1.46e-09.

Additionally, the analysis revealed that 29 genes were associated with the term "Cellular process" (GO:0009987), including *g.gyrB, g.gyrA, g.rpsF, g.rplI, g.aac, g.iniA, g.murB, g.mshA, g.rplK, g.rplA, g.rpoB, g.rpsL, g.tap, g.embR, g.murA, g.rpsA, g.tlyA, g.katG, g.pncA, g.blaC, g.kasA, g.folC, g.thyA, g.efpA, g.embC, g.embA, g.embB, g.ethA*, and *g.gid*. The strength of this enrichment was 0.42, with an FDR of 1.14e-07. Other significant associations were found for terms such as "Response to stimulus" (GO:0050896), "Cellular component organization or biogenesis" (GO:0071840), "Cellular metabolic process" (GO:0044237), and "Cellular macromolecule biosynthetic process" (GO:0034645), among others (see Table 4.8).

The MDR genes demonstrated enrichment in RNA polymerase, specifically involving the genes *g.rpoB, g.rpoC, g.rpoZ*, and *g.rpoA* Table 4.9. Furthermore, the functional term GO:0006351 - Transcription, DNA-templated was enriched in the same set of genes (*g.rpoB, g.rpoC, g.rpoZ*, and *g.rpoA*). Similarly, the term GO:0044260 - Cellular macromolecule metabolic process showed enrichment in multiple genes, including *g.msrA, g.rpoB, g.rpoC, g.fusA1, g.rpoZ, g.inhA, g.msrB*, and *g.rpoA*.

Additionally, the term GO:0034645 - Cellular macromolecule biosynthetic process was enriched in *g.rpoB, g.rpoC, g.fusA1, g.rpoZ, g.inhA*, and *g.rpoA*. Another term, GO:0010467 - Gene expression, exhibited enrichment in the genes *g.rpoB, g.rpoC, g.fusA1, g.rpoZ*, and *g.rpoA*. Moreover, the term GO:0003899 - DNA-directed 5-3 RNA polymerase activity was associated with the genes *g.rpoB, g.rpoC, g.rpoZ*, and *g.rpoA*. Additionally, the term GO:0008113 - Peptide-methionine (S)-S-oxide reductase activity was enriched in the genes *g.msrA* and *g.msrB*. Furthermore,

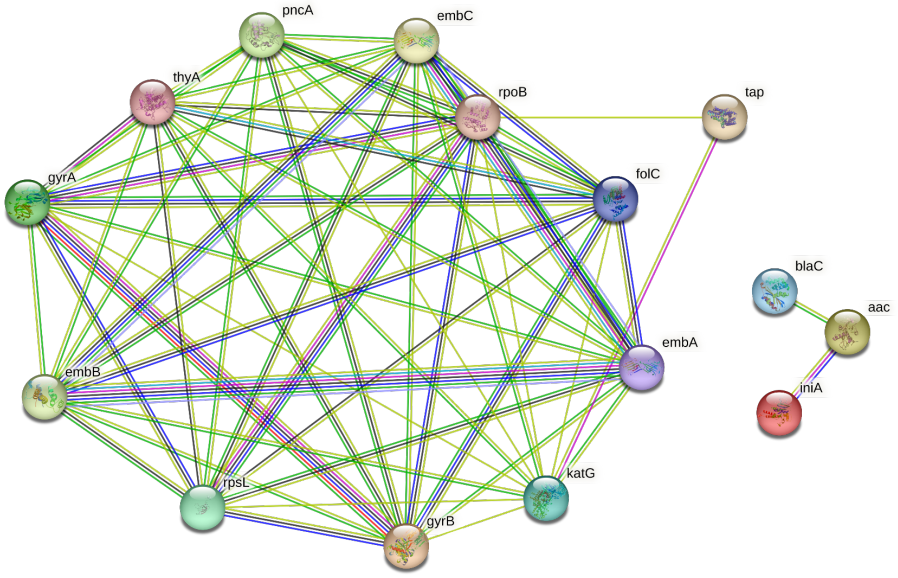


FIGURE 4.8: Network analysis of the genes responsible for AMR pathway for the common 15 genes between MDR and XDR.

the terms HSA-9639775 - Antimicrobial action and antimicrobial resistance in *M. tuberculosis* and HSA-9635486 - Infection with Mycobacterium TB showed enrichment in the genes *g.rpoB, g.rpoC, g.rpoZ*, and *g.rpoA*, while the latter term also included *g.msrA*. Through the utilization of STRING analysis, an examination of the common genes shared by MDR and XDR strains revealed distinct clusters of interactions (see Table 4.8 and 4.9). Specifically, the genes *blaC, aac*, and *iniA* formed a separate cluster, wherein blaC interacted with aac, aac interacted with *iniA*, but no direct interaction was observed between *blaC* and *iniA*. This suggests an indirect interaction between *blaC* and *iniA* through the involvement of the aac gene.

The enrichment analysis aimed to identify enriched pathways and functions associated with these variants, with a focus on differentiating between MDR and XDR strains of *M. tuberculosis*. Two distinct analyses were performed: one targeting the common genes shared between MDR and XDR strains, and the other specifically examining the unique genes associated with MDR strains. In the analysis of common novel variant genes, a significant enrichment signal was observed for the Antimicrobial resistance pathway. Fifteen genes, including *gyrB, gyrA, aac, iniA, tap, rpoB, rpsL, katG, pncA, blaC, floC, thyA, embC, embA, and embB*, exhibited

Table 4.7: Enriched pathways for the MDR novel genes in TB.

Term ID	**Term description**	**Observed gene count**	**Strength**	**False discovery rate**	**Genes**
mtv03020	RNA polymerase	4	2.51	3.26E-07	rpoB, rpoC, rpoZ, rpoA
GO:0006351	Transcription, DNA-templated	4	1.85	0.00055	*rpoB, rpoC, rpoZ, rpoA*
GO:0044260	Cellular macromolecule metabolic process	8	0.89	0.00055	*msrA, rpoB, rpoC, fusA1, rpoZ, inhA, msrB, rpoA*
GO:0034645	Cellular macromolecule biosynthetic process	6	1.02	0.0019	*rpoB, rpoC, fusA1, rpoZ, inhA, rpoA*
GO:0010467	Gene expression	5	1.05	0.0083	*rpoB, rpoC, fusA1, rpoZ, rpoA*
GO:0003899	DNA-directed 5-3 RNA polymerase activity	4	2.43	3.14E-06	*rpoB, rpoC, rpoZ, rpoA*
GO:0008113	Peptide-methionine (S)-S-oxide reductase activity	2	2.6	0.0052	*msrA, msrB*
HSA-9639775	Antimicrobial action and antimicrobial resistance in *M. tuberculosis*	4	2.6	3.71E-08	*rpoB, rpoC, rpoZ, rpoA*
HSA-9635486	Infection with *M. tuberculosis*	5	1.59	1.57E-06	*msrA, rpoB, rpoC, rpoZ, rpoA*

Table 4.8: Enriched pathways for the MDR and XDR common genes in TB

Term ID	Term Description	Gene Count	Strength	FDR	Genes
GO:0034335	DNA negative supercoiling activity	2	2.09	0.0335	*gyrB, gyrA*
GO:0052636	Arabinosyltran-sferase activity	3	1.96	0.0069	*embC,embA,embB*
GO:0006265	DNA topological change	2	1.91	0.0306	*gyrB,gyrA*
GO:0017001	Antibiotic catabolic process	2	1.91	0.0306	*aac,blaC*
KW-0799	Topoisomerase	2	1.91	0.0228	*gyrB,gyrA*
GO:0017144	Drug metabolic process	4	1.73	0.00042	*aac,pncA,blaC, ethA*
KW-0046	Antibiotic resistance	15	1.6	1.53E-17	*gyrB,gyrA,aac, iniA,rpoB,rpsL, tap,katG,pncA, blaC,folC,thyA, embC,embA,embB*
GO:0046677	Response to antibiotic	14	1.45	1.61E-13	*gyrB,gyrA,aac, iniA,rpoB,rpsL, tap,pncA,blaC, folC,thyA,embC, embA,embB*
GO:0071949	FAD binding	3	1.36	0.0335	*murB,Rv1260, ethA*
GO:0006364	rRNA processing	3	1.31	0.0295	*rplA,tlyA,gid*
GO:0042254	Ribosome biogenesis	4	1.18	0.0112	*rplK,rplA,tlyA, gid*
GO:0019843	rRNA binding	5	1.17	0.0069	*rpsF,rplI,rplK, rplA,rpsL*

Term ID	Term Description	Gene Count	Strength	FDR	Genes
KW-0699	rRNA-binding	5	1.17	0.0014	*rpsF,rplI,rplK, rplA,rpsL*
GO:0003735	Structural constituent of ribosome	6	1.11	0.0069	*rpsF,rplI,rplK, rplA,rpsL,rpsA*
KW-0689	Ribosomal protein	6	1.11	0.00074	*rpsF,rplI,rplK, rplA,rpsL,rpsA*
mtv03010	Ribosome	6	1.09	0.0016	*rpsF,rplI,rplK, rplA,rpsL,rpsA*
GO:0071555	Cell wall organization	5	1.08	0.0057	*murB,murA,embC, embA,embB*
KW-0961	Cell wall biogenesis/degradation	5	1.07	0.0034	*murB,murA,embC, embA,embB*
KW-0694	RNA-binding	7	1.05	0.00041	*rpsF,rplI,rplK, rplA,rpsL,rpsA, tlyA*
GO:0042221	Response to chemical	15	1.04	1.46E-09	*gyrB,gyrA,aac, iniA,rpoB,rpsL, tap,katG,pncA, blaC,folC,thyA, embC,embA,embB*
KW-0328	Glycosyltransferase	4	0.97	0.0312	*mshA,embC,embA, embB*
GO:0044085	Cellular component biogenesis	9	0.89	0.00039	*murB,rplK,rplA, murA,tlyA,embC, embA,embB,gid*
GO:0009273	Peptidoglycan-based cell wall biogenesis	5	0.89	0.0261	*murB, murA, embC, embA, embB*
GO:0003723	RNA binding	7	0.89	0.0069	*rpsF, rplI, rplK, rplA, rpsL, rpsA, tlyA*
GO:0006412	Translation	6	0.86	0.0112	*rpsF, rplI, rplK, rplA, rpsL, rpsA*
GO:0043603	Cellular amide metabolic process	9	0.84	0.00069	*rpsF, rplI, rplK, rplA, rpsL, rpsA, pncA, blaC, folC*

Term ID	Term Description	Gene Count	Strength	FDR	Genes
GO:0071840	Cellular component organization or biogenesis	11	0.83	0.00012	*gyrB, gyrA, murB, rplK, rplA, murA, tlyA, embC, embA, embB, gid*
GO:0016043	Cellular component organization	8	0.8	0.0026	*gyrB, gyrA, murB, rplK, murA, embC, embA, embB*
GO:0043604	Amide biosynthetic process	7	0.8	0.0078	*rpsF, rplI, rplK, rplA, rpsL, rpsA, folC*
GO:0010467	Gene expression	9	0.79	0.0013	*rpsF, rplI, rplK, rplA, rpoB, rpsL, rpsA, tlyA, gid*
GO:0034645	Cellular macromolecule biosynthetic process	11	0.76	0.00039	*gyrB,gyrA,rpsF, rplI,murB,rplK, rplA,rpoB,rpsL, murA,rpsA*
GO:0050896	Response to stimulus	16	0.69	1.19E-05	*gyrB, gyrA, aac, iniA, rpoB, rpsL, tap, embR, katG, pncA, blaC, folC, thyA, embC, embA, embB*
GO:0003676	Nucleic acid binding	11	0.61	0.0069	*gyrB, gyrA, rpsF, rplI, rplK, rplA, rpoB, rpsL, embR, rpsA, tlyA*
GO:0044271	Cellular nitrogen compound biosynthetic process	11	0.6	0.0041	*rpsF, rplI, rplK, rplA, rpoB, rpsL, murA, rpsA, pncA, folC, thyA*
GO:0044260	Cellular macromolecule metabolic process	13	0.58	0.0013	*gyrB, gyrA, rpsF, rplI, murB, rplK, rplA, rpoB, rpsL, murA, rpsA, tlyA, gid*

Term ID	Term Description	Gene Count	Strength	FDR	Genes
GO:0034641	Cellular nitrogen compound metabolic process	16	0.52	0.00069	*gyrB,gyrA,rpsF, rplI,rplK,rplA, rpoB,rpsL,murA, rpsA,tlyA,pncA, blaC,folC,thyA, gid*
GO:1901566	Organonitrogen compound biosynthetic process	11	0.52	0.0146	*rpsF,rplI,murB, rplK,rplA,rpsL, murA,rpsA,pncA, folC,thyA*
GO:0044249	Cellular biosynthetic process	17	0.49	0.00068	*gyrB,gyrA,rpsF, rplI,murB,mshA, rplK,rplA,rpoB, rpsL,murA,rpsA, pncA,kasA,folC, thyA,embC*
GO:1901576	Organic substance biosynthetic process	17	0.48	0.00083	*gyrB,gyrA,rpsF, rplI,murB,mshA, rplK,rplA,rpoB, rpsL,murA,rpsA, pncA,kasA,folC, thyA,embC*
GO:0016740	Transferase activity	11	0.47	0.0396	*aac, mshA, rpoB, murA, tlyA, kasA, thyA, embC, embA, embB, gid*
KW-0808	Transferase	11	0.45	0.0325	*aac, mshA, rpoB, murA, tlyA, kasA, thyA, embC, embA, embB, gid*
GO:0009987	Cellular process	29	0.42	1.14E-07	*gyrB, gyrA, rpsF, rplI, aac, iniA, murB, mshA, rplK, rplA, rpoB, rpsL, tap, embR, murA, rpsA, tlyA, katG, pncA, blaC, kasA, folC, thyA, efpA, embC, embA, embB, ethA, gid*
GO:0006807	Nitrogen compound metabolic process	17	0.4	0.006	*gyrB, gyrA, rpsF, rplI, murB, rplK, rplA, rpoB, rpsL, murA, rpsA, tlyA, pncA, blaC, folC, thyA, gid*

Term ID	Term Description	Gene Count	Strength	FDR	Genes
GO:0097159	Organic cyclic compound binding	16	0.4	0.0204	*gyrB, gyrA, rpsF, rplI, murB, rplK, rplA, rpoB, rpsL, Rv1260, embR, rpsA, tlyA, katG, folC, ethA*
GO:1901363	Heterocyclic compound binding	16	0.4	0.0204	*gyrB,gyrA,rpsF, rplI,murB,rplK, rplA,rpoB,rpsL, Rv1260,embR,rpsA, tlyA,katG,folC, ethA*
GO:0044237	Cellular metabolic process	23	0.39	0.0003	*gyrB,gyrA,rpsF, rplI,aac,murB, mshA,rplK,rplA, rpoB,rpsL,murA, rpsA,tlyA,katG, pncA,blaC,kasA, folC,thyA,embC, ethA,gid*
GO:0071704	Organic substance metabolic process	21	0.35	0.0025	*gyrB,gyrA,rpsF, rplI,aac,murB, mshA,rplK,rplA, rpoB,rpsL,murA, rpsA,tlyA,pncA, blaC,kasA,folC, thyA,embC,gid*
GO:0005488	Binding	19	0.3	0.0369	*gyrB,gyrA,rpsF, rplI,murB,mshA, rplK,rplA,rpoB, rpsL,Rv1260,embR, rpsA,tlyA,katG, pncA,kasA,folC, ethA*

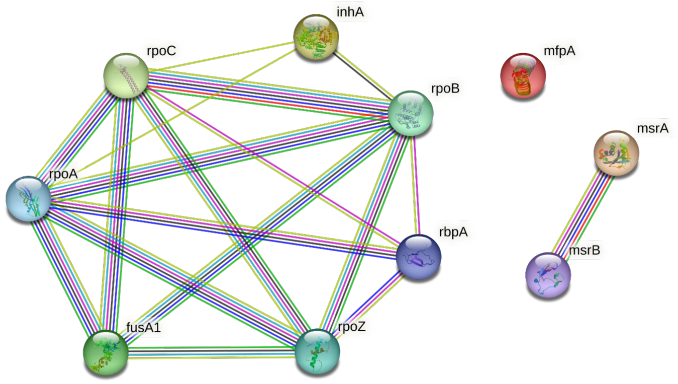


FIGURE 4.9: Network analysis of the genes responsible for AMR pathway for the common genes MDR.

strong enrichment in this pathway, with a very low false discovery rate (FDR). Furthermore, several other functional terms were identified through enrichment analysis. Genes involved in RNA-binding, ribosomal proteins, rRNA-binding, cell wall biogenesis/degradation, topoisomerase activity, glycosyltransferase activity, and transferase activity were found to be significantly associated with the common novel variant genes. Gene Ontology analysis revealed additional insights into the functional annotations of the genes.

Notably, genes associated with Nucleic acid binding, RNA binding, Structural constituent of ribosome, and rRNA binding exhibited enrichment signals. Various other terms, including Arabinosyltransferase activity, Organic cyclic compound binding, Heterocyclic compound binding, DNA negative supercoiling activity, FAD binding, Binding, and Transferase activity, also showed significant associations. The enrichment analysis of Gene Ontology Biological Processes highlighted the involvement of the novel variant genes in specific biological processes. For instance, the term "Response to antibiotic" (GO:0046677) was significantly associated with 14 genes, while the term "Response to chemical" (GO:0042221) was associated with 15 genes.

In the MDR strains, specific genes related to RNA polymerase, such as rpoB, rpoC, rpoZ, and rpoA, demonstrated enrichment. These genes were also associated with terms like Transcription, DNA-templated and Cellular macromolecule metabolic process. Furthermore, the genes *msrA* and *msrB* were enriched in the

term Peptide-methionine (S)-S-oxide reductase activity. Enrichment was also observed for terms related to antimicrobial action and resistance in *M. tuberculosis*, including HSA-9639775 (Antimicrobial action) and HSA-9635486 (Infection with *M. tuberculosis*).

These results provide valuable insights into the functional characteristics of the common novel variant genes, highlighting their involvement in nucleic acid binding, RNA binding, ribosome structure, and other important molecular activities. These associations provide insights into the functional roles of the identified genes in various cellular processes and metabolic pathways. These results provide valuable insights into the potential molecular mechanisms and pathways involved in the development of antibiotic resistance and cellular processes related to ribosomal functions, RNA-binding, cell wall biogenesis/degradation, and enzymatic activities. These findings suggest that these genes play crucial roles in conferring resistance to antibiotics.

A total of 15 genes were involved in antibiotic resistance pathways. These 15 genes were categorized based on gene families, and their functions, including *aac, gyrB, and gyrA*, are antibiotic resistance-related genes. Aac is involved in antibiotics to modify aminoglycoside, while *gyrA* and *gyrB* encode the A and B subunits of DNA gyrase, respectively, and confer antibiotic resistance [256]. It is reported that *rpoB, rpsL*, and *iniA* are drug resistance-related genes in which iniA isoniazid is the resistance in *M. tuberculosis* [257]. The rpoB synthesizes the B subunit of RNA polymerase and is involved in the resistance of rifampicin, an important antibiotic in TB. Rpsl encodes the ribosomal protein S12 and is involved in streptomycin resistance [258]. The cell wall-related genes *embC, embA*, and *embB* are major components of the mycobacterial cell wall involved in the biosynthesis of arabinogalactans [202]. Furthermore, the remaining 6 genes were categorized based on their function in which the tap gene was involved in synthesizing antigenic peptide transporter, which is involved in antigens of the immune system [259]. *katG* is used to activate the drug isoniazid, which is anti- [260]. *pncA* is involved in the pyrazinamide drug activation in anti-TB [232]. In contrast, *blaC* is involved in

resistance to B lactam antibiotics, and *folC* and *thyA* are folate metabolism and DNA synthesis enzymes, respectively [261, 262].

In addition, the remaining 16 genes were not involved in the antimicrobial pathway in our study. *embR* is also a component of the mycobacterial cell wall. It is involved in the biosynthesis of arabinogalactan by encoding a transcriptional regulator [263]. *kasA* synthesized the beta ketoacyl ACP synthase involved in mycolic acid biosynthesis, an important component of the mycobacterial cell wall [264]. The *mshA* gene synthesizes mycothiol, an important thiol oxidant in mycobacterium species [265]. The rpsA gene encodes the S1 protein of the ribosome, which plays an important role in translation and ribosome assembly [266]. The gidB codes a tRNA uracil methyltransferase involved in the molecules modification of *tRNA* [267]. The *iniC* gene is involved in the resistivity against isoniazid in *M. tuberculosis* [265], while ethA encodes an ethA enzyme that acts in activating the antituberculosis prodrug ethionamide [268]. The Rv1258c is involved in multidrug efflux pump in *M. tuberculosis* against isoniazid, Streptomycin, and pyrazinamide but not to other drugs in *M. tuberculosis* [269].

The tet gene allows resistance against tetracycline by decreasing its accumulation in the cell wall [246]. The erm gene acts to produce dimethylation at adenine residue at position 2085 of 23S rRNA and minimize the affinity between ribosomes and antibiotics [252]. The *murA* gene is responsible for cell wall formation by adding enolpyruvyl to UDP-N-acetylglucosamine [216]. The Planobispora-rosea-EF-Tu (tuf) is involved in GTPase activity and GTP binding [251]. The aac gene involved in the catalysis of coenzyme-dependent acetylation of 2' hydroxyl of aminoglycosides [206]. The ribD gene performs diamino-hydroxy-phospho-ribosyl-amino pyrimidine deaminase activity [218].

The *rpoC* gene encodes a subunit of RNA polymerase, and changes in amino acid in rpoC alter RNA polymerase's structure and function, decreasing the binding affinity with rifampicin and is therefore involved in drug resistance [270]. Many studies revealed that mutations in *rpoC* is associated with enhanced in vitro fitness and were identified in higher proportion among MDR-TB isolates from countries

with significant MDR-TB burdens [271]. The researchers identified mutations at codon G332R in rpoC, which have high protein stability, lower flexibility, and favorable compensatory effects reported in Khyber Pakhtunkhwa [272] and South Africa isolates [271].

Eight different mutations in *rpoC*, including G332R, F452C, D485Y, V483A, V483G, I491T, Q523E, and H525Q, were identified, which is previously reported as a phylogenetic marker for the Latin American Mediterranean family of *M. tuberculosis* [273].

Researchers reported more frequent mutation at codon 452 (F452L) in rpoC for the samples of South Korean patients [274]. In our study, mutations at codon 864, 887, and 898 (V864I, Q887K, and A898T) were identified, which was not reported in the previous studies. Hence, these could be our potential targets for controlling antimicrobial resistivity in *M. tuberculosis*.

The study revealed that MDR strains of *M. tuberculosis* exhibited enrichment in genes related to RNA polymerase and peptide-methionine (S)-S-oxide reductase activity, as well as terms associated with antimicrobial action and resistance Furthermore, 15 genes participated in antibiotic resistance pathways, including aminoglycoside modification, DNA gyrase, isoniazid resistance, ribosomal proteins, and cell wall biosynthesis. The remaining 16 genes had diverse functions unrelated to the antimicrobial pathway, such as cell wall formation, tRNA modification, multidrug efflux pump, and enzyme activities.

The study also highlighted specific mutations in rpoC linked to enhanced fitness and drug resistance. Genes such as rpoB,rpoC,rpoZ,rpoA which have shown relevance to *M. tuberculosis* drug resistance and have gained novel variants, should be studied further. These findings provide valuable insights into the genetic variants and pathways associated with MDR and XDR drug resistance in *M. tuberculosis*, potentially serving as targets for controlling antimicrobial resistivity in tuberculosis. Further investigations are necessary to fully comprehend the mechanisms and implications of these novel variants in drug resistance.

TABLE 4.9: Hub genes involved in MDR and XDR.

Rank	Name	Score
1	*KatG*	20
1	*RpoB*	20
3	*RpsL*	18
3	*PncA*	18
3	*ThyA*	18

4.5 Hub Gene Identification

The identification of the top five hub genes, which exhibit the highest number of interactions with other genes in the entire network, was performed using Cytoscape and its module CytoHubba. This approach allowed for the identification of hub genes based on degrees of connectivity. Remarkably, the analysis revealed that the genes *rpoB* and *katG* displayed the most extensive interactions within this network (Figure 4.3), indicating their significant involvement and functional mobility with other genes. The remaining three out of the five hub genes were identified as *rpsL, thyA*, and *pncA* (Table 4.5).

The genes associated with both XDR and MDR strains - roles in regulating biological processes associated with the analyzed condition or disease. A comprehensive network of interactions was observed among all other genes, forming an intricate map. Notably, each gene exhibited the capability to receive and transmit signals while interacting with other proteins.Rank indicates a gene's position in the list of hub genes, with the top-ranked genes having the highest connectivity to others in the network. Score represents a numerical measure of a gene's centrality or importance within the network, with higher scores indicating greater connectivity and significance. These parameters helped prioritize genes based on their network

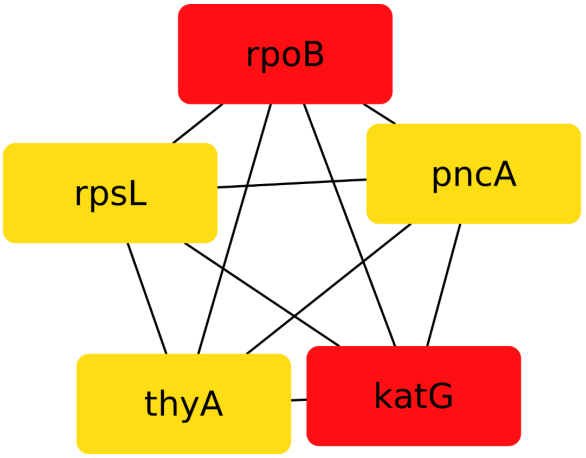


FIGURE 4.10: Top 5 hub genes among the genes that show the most AMR functions.

interactions. The specific method for calculating scores for each of used for network analysis in CytoHubba. The *katG* and *rpoB* showed highest score i.e. 20 as compared to the other genes. This indicates that both of these genes plays highly significant role in resistance against the drugs categorized as MDR and XDR. All of these genes have also be revealed through genomic data analysis under common novel variant in Table 4.5.

The *katG* and *rpoB* showed highest score i.e. 20 as compared to the other genes. This indicates that both of these genes plays highly significant role in resistance against the drugs categorized as MDR and XDR.All of these genes have also be revealed through genomic data analysis under common novel variant in Table 4.5.

In association rule mining, it was found that the highest occurrences of drug resistance were associated with the genes *tlyA, gyrA, rrs, rpsl, inhA, rpoB*, and

katG. Notably, four out of the five identification hub genes are part of this list. Consequently, the phenotypic results showa a similarity of 80% when compared to the genotype-based results. The observed high similarity between phenotypic and genotype-based results underscores the reliability of genetic information in predicting drug resistance.

Chapter 5

Conclusions and Future Directions

5.1 Conclusions

5.1.1 Understanding Demographic Characteristics of TB Patients and Their Association with Treatment Outcomes, Lineage, and Drug Resistance

The significance of this objective was to identify and assess cases of drug-resistant TB (DR-TB) based on various demographic factors, including resistance type, age at onset, gender, and treatment results. A dataset was subjected to pattern recognition techniques to pinpoint important features. Resistance types, such as pre-extensively drug-resistant (pre-XDR), extensively drug-resistant (XDR), sensitive, single-drug resistant (mono DR), multi-drug resistant (poly DR), and MDR non-XDR, were visually represented in relation to different treatment outcomes using bar charts. The analysis unveiled that XDR and MDR non-XDR were the most prevalent forms of TB resistance and were linked to unfavorable outcomes like death or treatment failure. Age of onset analysis, as depicted in box plots, revealed that the median age for both resistance types hovered around 40, with a

slight variation between XDR and MDR non-XDR TB cases.Gender wise analysis of drug resistant isolates revealed that gender has no specific impact on drug resistance.Lineage relation with respect to drug resistance predicted that the Beijing strain showed maximum resistance to antobiotics. The significance of this work lies in its multifaceted approach to risk factor analysis within DR-TB. By uncovering these relationships, the research can contribute to a the broader epidemiological understanding of DR-TB. These insights will undoubtedly aid in the development of targeted interventions and policies to mitigate the risks associated with MDR-TB.

5.1.2 To Uncover Significant MDR and XDR Patterns in Antimicrobial Susceptibility Testing Data with Association Rule Mining

This objective was addressed by employing association rule mining to uncover significant drug resistance patterns in TB treatment. Filtering through the TB portals dataset of generated rules, meaningful associations were identified and categorized based on the drugs involved and the types of resistance. The minimum support was 0.01 and the confidence was 0.9. The association rule with maximum support of 0.5 was [RIF to INH] which represents a strong association between isoniazid (INH) and rifampicin (RIF) resistance, reinforced by genetic links between INH target genes (*katG* and *inhA*) and RIF resistance mutations in *rpoB*.

5.1.3 To Analyze Genomic Sequences of Drug-Resistant TB Isolates and Identify Existing and Unique Mutations to Establish Relationship with Pattern Through Data-mining.

This study utilized the ARIBA pipeline to analyze MDR and XDR strains of TB, focusing on identifying and validating variants associated with drug resistance in

the strains available at TB portals. The research provided valuable insights into the genetic variants contributing to the development of resistance in these TB strains. A total of 27 frequently occurring common variants in the data were observed, including genes such as *thyA, embR, embB, embC, folC, kasA, aac, mshA, iniA, rpsA, gyrA, gidB, iniC, pncA, ethA, Rv1258c, embA, murA, iniB, ribD, gyrB, tlyA, efpA, rpsL, katG, rpoB*, and *blaC*. These genes were associated with both MDR and XDR TB.

Additionally, 12 other genes were characterized as novel variants conferring resistance to MDR TB or XDR TB. Interestingly, a literature survey revealed that these genes had not been previously reported for antimicrobial resistance in TB. However, they have been associated with resistance mechanisms in other bacterial species. The presence of these sequences in TB isolates from TB portals suggests their potential involvement in resistance mechanisms in *M. tuberculosis*. The discovery of common genes and novel variants associated with multi-drug resistance (MDR) and extremely drug-resistant (XDR) TB enhances our ability to diagnose and treat these challenging cases effectively. This newfound knowledge has the potential to inform the development of more targeted and efficient diagnostic tools and therapies, ultimately aiding in the global fight against drug-resistant TB.

5.1.4 To Perform Functional Enrichment Analysis to Understand Gene Functions and Assess the Functional Significance in TB Drug Resistance.

The enrichment analysis conducted in this study aimed to identify significant pathways and functions associated with genetic variants in MDR and XDR strains of *M. tuberculosis*. Two distinct analyses were performed, one focusing on common genes shared between MDR and XDR strains, and the other examining unique genes associated with MDR strains. The analysis of common novel variant genes revealed a strong enrichment signal for the antimicrobial resistance pathway. the term "Response to antibiotic" (GO:0046677) was significantly associated with 14

genes, while the term "Response to chemical" (GO:0042221) was associated with 15 genes. These genes, namely *gyrB, gyrA, aac, iniA, tap, rpoB, rpsL, katG, pncA, blaC, floC, thyA, embC, embA*, and *embB*, exhibited a strong enrichment signal and a very low false discovery rate Additionally, several functional terms related to RNA-binding, ribosomal proteins, cell wall biogenesis, glycosyltransferase activity, and more were identified. Gene Ontology analysis highlighted the involvement of novel variant genes in biological processes such as "Response to antibiotic" and "Response to chemical."

In MDR strains, specific genes related to RNA polymerase demonstrated enrichment, along with terms related to transcription and cellular macromolecule metabolic processes. Additionally, terms related to antimicrobial action and resistance in *M. tuberculosis* were observed.

For future directions, further research is needed to explore the functional implications of the identified genes and variants in drug resistance mechanisms. Understanding the specific roles of these genes in different pathways and processes will contribute to the development of more effective strategies to combat antibiotic resistance in *M. tuberculosis*.

5.1.5 To Identify Hub Genes and Construct a Network to Understand Molecular Interactions Related to TB Drug Resistance.

The network analysis using Cytoscape and CytoHubba identified the top five hub genes with the highest degrees of connectivity within the network. Notably, *rpoB* and *katG* emerged as the most extensively interconnected genes, highlighting their crucial roles and functional interactions with other genes.

The remaining three hub genes, namely *rpsL, thyA*, and *pncA*, also exhibited significant connectivity within the network. These findings provide valuable insights

into the key genes that play pivotal roles in the genetic interactions associated with the studied biological process.

The research emphasizes the need to uncover hidden pathways and mechanisms underlying TB drug resistance. Further investigations should aim to identify novel genetic variants and genes that may contribute to resistance, potentially unveiling new targets for intervention. Lineage base genomic patterns must be analyzed to enhance TB treatment efficacy, personalized medicine approaches should be explored. Understanding how individual genetic variations, including those in hub genes like *rpoB* and *katG*, impact treatment responses can guide the development of tailored therapies for DR-TB patients.

Given the genetic diversity of TB strains in different regions, tailoring treatment regimens based on regional genetic patterns is essential. Future research should delve into region-specific genetic variations and their implications for treatment outcomes. Novel variants identified can be subject to detailed analysis to explore their role at molecular level in *M. tuberculosis*. Developing generic drug designs capable of targeting multiple genes associated with TB drug resistance is crucial. Investigating how different resistance mechanisms interact and exploring the potential for broad-spectrum drugs can lead to more effective treatments.

5.2 Future Directions

5.2.1 Evaluation at Proteomic Level

To gain a comprehensive understanding of resistance mechanisms, it's essential to evaluate them at the proteomic level. This entails evaluating the proteins involved in resistance pathways to gain a deeper understanding of how they contribute to drug resistance. By examining changes in protein expression, modifications, and interactions, researchers can uncover valuable insights into the underlying mechanisms driving resistance. This proteomic approach allows for a comprehensive

analysis of the molecular landscape associated with resistance, providing a foundation for the development of targeted therapies. Ultimately, by delving into the proteomic level of resistance, researchers can unravel the complexities of drug resistance and pave the way for more effective treatment strategies.

5.2.2 Pathway Analysis

Pathways involved in resistance must be targeted to understand the mechanism of resistance and the area of drug target . To gain a comprehensive understanding of resistance mechanisms and identify potential drug targets, it is imperative to uncover the molecular mechanisms involved in resistance pathways. This exploration can reveal critical points of vulnerability within these pathways, which may serve as promising targets for therapeutic intervention., paving the way for the development of more effective treatments against resistant infections.

5.2.3 Research on Model Organisms

Implication at Animal Model for Pharmacokinetic Studies to understand the drug behavior after mutations. Implementation of findings from bioinformatics analysis to animal models to bridge the gap between computational predictions and experimental validation.

5.2.4 Genomic Epidemiology

Utilize genomic epidemiology approaches to elucidate transmission dynamics and population structure of drug-resistant Mycobacterium tuberculosis strains, informing targeted control measures and surveillance strategies.

5.2.5 Precision Medicine

The understanding of drug-resistant TB is essential for the development and implementation of precision medicine approaches that optimize treatment outcomes, minimize the spread of resistance, and ultimately accelerate progress towards TB elimination.

5.2.6 One Health Approach

Adopt a One Health approach by integrating data from animal models, human clinical studies, and environmental surveillance to understand the interconnectedness of drug-resistant TB across different host species and environments. Identify common risk factors, transmission routes, and intervention strategies to control TB transmission and drug resistance emergence.

5.2.7 Real-Time Surveillance

The information derived through this research can be utilize to develop a system that utilizes bioinformatics tools to monitor and track emerging drug resistance mutations and patterns in Mycobacterium tuberculosis populations, facilitating timely intervention and management strategies.

5.2.8 More Data Integration

The dataset should be expanded by incorporating genomic, clinical, and epidemiological data from diverse geographic regions and populations to capture a broader spectrum of drug resistance patterns in Mycobacterium tuberculosis.

Bibliography

[1] W. H. Organization, "Tuberculosis and diabetes: invest for impact: information note," 2023.

[2] K. A. Alexander, P. N. Laver, A. L. Michel, M. Williams, P. D. van Helden, R. M. Warren, and N. C. G. van Pittius, "Novel mycobacterium tuberculosis complex pathogen, m. mungi," *Emerging infectious diseases*, vol. 16, no. 8, p. 1296, 2010.

[3] R. G. Barletta and D. J. Steffen, "Mycobacteria," *Veterinary Microbiology*, pp. 345–359, 2022.

[4] J.-H. Park, D. Shim, K. E. S. Kim, W. Lee, and S. J. Shin, "Understanding metabolic regulation between host and pathogens: New opportunities for the development of improved therapeutic strategies against mycobacterium tuberculosis infection," *Frontiers in Cellular and Infection Microbiology*, vol. 11, p. 635335, 2021.

[5] M. De Martino, L. Lodi, L. Galli, and E. Chiappini, "Immune response to mycobacterium tuberculosis: a narrative review," *Frontiers in pediatrics*, vol. 7, p. 350, 2019.

[6] J. Advani, R. Verma, O. Chatterjee, P. K. Pachouri, P. Upadhyay, R. Singh, J. Yadav, F. Naaz, R. Ravikumar, S. Buggi, *et al.*, "Whole genome sequencing of mycobacterium tuberculosis clinical isolates from india reveals genetic heterogeneity and region-specific variations that might affect drug susceptibility," *Frontiers in microbiology*, vol. 10, p. 309, 2019.

[7] R. D. Kanabalan, L. J. Lee, T. Y. Lee, P. P. Chong, L. Hassan, R. Ismail, and V. K. Chin, "Human tuberculosis and mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery," *Microbiological Research*, vol. 246, p. 126674, 2021.

[8] S. Bagcchi, "Who's global tuberculosis report 2022," *The Lancet Microbe*, vol. 4, no. 1, p. e20, 2023.

[9] M. Kumar, S. Jaiswal, K. K. Sodhi, P. Shree, D. K. Singh, P. K. Agrawal, and P. Shukla, "Antibiotics bioremediation: Perspectives on its ecotoxicity and resistance," *Environment International*, vol. 124, pp. 448–461, 2019.

[10] Y. Habboush and N. Guzman, "Antibiotic resistance," 2018.

[11] R. Urban-Chmiel, A. Marek, D. Stkepien-Pysniak, K. Wieczorek, M. Dec, A. Nowaczek, and J. Osek, "Antibiotic resistance in bacteria—a review," *Antibiotics*, vol. 11, no. 8, p. 1079, 2022.

[12] L. Nguyen, "Antibiotic resistance mechanisms in m. tuberculosis: an update," *Archives of toxicology*, vol. 90, pp. 1585–1604, 2016.

[13] K. Moopanar, A. N. G. Nyide, S. Senzani, and N. E. Mvubu, "Clinical strains of mycobacterium tuberculosis exhibit differential lipid metabolism-associated transcriptome changes in in vitro cholesterol and infection models," *Pathogens and Disease*, vol. 81, 2023.

[14] Y. Liu, M. Matsumoto, H. Ishida, K. Ohguro, M. Yoshitake, R. Gupta, L. Geiter, and J. Hafkin, "Delamanid: from discovery to its use for pulmonary multidrug-resistant tuberculosis (mdr-tb)," *Tuberculosis*, vol. 111, pp. 20–30, 2018.

[15] P. Ajawatanawong, H. Yanai, N. Smittipat, A. Disratthakit, N. Yamada, R. Miyahara, S. Nedsuwan, W. Imasanguan, P. Kantipong, B. Chaiyasirinroje, *et al.*, "A novel ancestral beijing sublineage of mycobacterium tuberculosis suggests the transition site to modern beijing sublineages," *Scientific reports*, vol. 9, no. 1, p. 13718, 2019.

[16] P. Miotto, Y. Zhang, D. M. Cirillo, and W. C. Yam, "Drug resistance mechanisms and drug susceptibility testing for tuberculosis," *Respirology*, vol. 23, no. 12, pp. 1098–1113, 2018.

[17] G. Xu, H. Liu, X. Jia, X. Wang, and P. Xu, "Mechanisms and detection methods of mycobacterium tuberculosis rifampicin resistance: The phenomenon of drug resistance is complex," *Tuberculosis*, vol. 128, p. 102083, 2021.

[18] P. Khumwan, S. Pengpanich, J. Kampeera, W. Kamsong, C. Karuwan, A. Sappat, P. Srilohasin, A. Chaiprasert, A. Tuantranont, and W. Kiatpathomchai, "Identification of s315t mutation in katg gene using probe-free exclusive mismatch primers for a rapid diagnosis of isoniazid-resistant mycobacterium tuberculosis by real-time loop-mediated isothermal amplification," *Microchemical Journal*, vol. 175, p. 107108, 2022.

[19] J. Espinosa-Pereiro, A. Sanchez-Montalva, M. L. Aznar, and M. Espiau, "Mdr tuberculosis treatment," *Medicina*, vol. 58, no. 2, p. 188, 2022.

[20] A. Gill, I. Ugalde, C. A. Febres-Aldana, and C. Tuda, "Fluoroquinolone resistant tuberculosis: A case report and literature review," *Respiratory Medicine Case Reports*, vol. 27, p. 100829, 2019.

[21] Y. Che, T. Yang, L. Lin, Y. Xiao, F. Jiang, Y. Chen, T. Chen, and J. Zhou, "Comparative utility of genetic determinants of drug resistance and phenotypic drug susceptibility profiling in predicting clinical outcomes in patients with multidrug-resistant mycobacterium tuberculosis," *Frontiers in Public Health*, vol. 9, p. 663974, 2021.

[22] A. N. Phyu, S. T. Aung, P. Palittapongarnpim, K. K. K. Htet, S. Mahasirimongkol, H. L. Aung, A. Chaiprasert, and V. Chongsuvivatwong, "Distribution of mycobacterium tuberculosis lineages and drug resistance in upper myanmar," *Tropical Medicine and Infectious Disease*, vol. 7, no. 12, p. 448, 2022.

[23] M. Humayun, J. Chirenda, W. Ye, I. Mukeredzi, H. A. Mujuru, and Z. Yang, "Effect of Gender on Clinical Presentation of Tuberculosis (TB) and Age-Specific Risk of TB, and TB-Human Immunodeficiency Virus Coinfection," *Open Forum Infectious Diseases*, vol. 9, p. ofac512, 10 2022.

[24] H. A. Taylor, D. W. Dowdy, A. R. Searle, A. L. Stennett, V. Dukhanin, A. A. Zwerling, and M. W. Merritt, "Disadvantage and the experience of treatment for multidrug-resistant tuberculosis (mdr-tb)," *SSM-Qualitative Research in Health*, vol. 2, p. 100042, 2022.

[25] M. A. Huaman and T. R. Sterling, "Treatment of latent tuberculosis infection—an update," *Clinics in chest medicine*, vol. 40, no. 4, pp. 839–848, 2019.

[26] C. O'Connor and M. F. Brady, "Isoniazid," no. PMID:32491549, 2022.

[27] V. A. Dartois and E. J. Rubin, "Anti-tuberculosis treatment strategies and drug development: challenges and priorities," *Nature Reviews Microbiology*, vol. 20, no. 11, pp. 685–701, 2022.

[28] D. Kalo, S. Kant, K. Srivastava, and A. K. Sharma, "Pattern of drug resistance of mycobacterium tuberculosis clinical isolates to first-line antituberculosis drugs in pulmonary cases," *Lung India: Official Organ of Indian Chest Society*, vol. 32, no. 4, p. 339, 2015.

[29] A. Minias, L. Zukowska, E. Lechowicz, F. Gkasior, A. Knast, S. Podlewska, D. Zygala, and J. Dziadek, "Early drug development and evaluation of putative antitubercular compounds in the-omics era," *Frontiers in Microbiology*, vol. 11, p. 618168, 2021.

[30] J. Gabrielsson, B. Meibohm, and D. Weiner, "Pattern recognition in pharmacokinetic data analysis," *The AAPS journal*, vol. 18, pp. 47–63, 2016.

[31] J. Cervantes, N. Yokobori, and B.-Y. Hong, "Genetic identification and drug-resistance characterization of mycobacterium tuberculosis using a portable sequencing device. a pilot study," *Antibiotics*, vol. 9, no. 9, p. 548, 2020.

[32] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," *International Journal on Computer Science and Engineering (IJCSE)*, vol. 2, no. 02, pp. 250–255, 2010.

[33] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using weka," *Bioinformatics*, vol. 20, no. 15, pp. 2479–2481, 2004.

[34] K. Raza, "Application of data mining in bioinformatics," *arXiv preprint arXiv:1205.1125*, 2012.

[35] V. Singh and K. Chibale, "Strategies to combat multi-drug resistance in tuberculosis," *Accounts of chemical research*, vol. 54, no. 10, pp. 2361–2376, 2021.

[36] Z. Yang, X. Zeng, and S. K.-W. Tsui, "Investigating function roles of hypothetical proteins encoded by the mycobacterium tuberculosis h37rv genome," *BMC genomics*, vol. 20, pp. 1–10, 2019.

[37] M. J. Nasiri, M. Zangiabadian, E. Arabpour, S. Amini, F. Khalili, R. Centis, L. D'Ambrosio, J. T. Denholm, H. S. Schaaf, M. van den Boom, *et al.*, "Delamanid-containing regimens and multidrug-resistant tuberculosis: A systematic review and meta-analysis," *International Journal of Infectious Diseases*, vol. 124, pp. S90–S103, 2022.

[38] Z. M. Tan, G. P. Lai, M. Pandey, T. Srichana, M. R. Pichika, B. Gorain, S. K. Bhattamishra, and H. Choudhury, "Novel approaches for the treatment of pulmonary tuberculosis," *Pharmaceutics*, vol. 12, no. 12, p. 1196, 2020.

[39] S. Y. Rodriguez-Takeuchi, M. E. Renjifo, and F. J. Medina, "Extrapulmonary tuberculosis: pathophysiology and imaging findings," *Radiographics*, vol. 39, no. 7, pp. 2023–2037, 2019.

[40] A. H. Baykan, H. S. Sayiner, E. Aydin, M. Koc, I. Inan, and S. M. Erturk, "Extrapulmonary tuberculosis: an old but resurgent problem," *Insights Into Imaging*, vol. 13, no. 1, pp. 1–21, 2022.

[41] C. Allix-Beguec, P. Supply, M. Wanlin, P. Bifani, and M. Fauville-Dufaux, "Standardised pcr-based molecular epidemiology of tuberculosis," *European Respiratory Journal*, vol. 31, no. 5, pp. 1077–1084, 2008.

[42] E. A. Nardell, "Transmission and institutional infection control of tuberculosis," *Cold Spring Harbor perspectives in medicine*, vol. 6, no. 2, p. a018192, 2016.

[43] A. Singh, R. Prasad, V. Balasubramanian, and N. Gupta, "Drug-resistant tuberculosis and hiv infection: current perspectives," *HIV/AIDS-Research and Palliative Care*, pp. 9–31, 2020.

[44] J. C. Palomino and A. Martin, "Drug resistance mechanisms in mycobacterium tuberculosis," *Antibiotics*, vol. 3, no. 3, pp. 317–340, 2014.

[45] R. L. Hunter, "The pathogenesis of tuberculosis–the koch phenomenon reinstated," *Pathogens*, vol. 9, no. 10, p. 813, 2020.

[46] J. A. Philips and J. D. Ernst, "Tuberculosis pathogenesis and immunity," *Annual Review of Pathology: Mechanisms of Disease*, vol. 7, pp. 353–384, 2012.

[47] V. Peddireddy, S. N. Doddam, and N. Ahmed, "Mycobacterial dormancy systems and host responses in tuberculosis," *Frontiers in immunology*, vol. 8, p. 84, 2017.

[48] T. Paulson, "Epidemiology: a mortal foe," *Nature*, vol. 502, no. 7470, pp. S2–S3, 2013.

[49] H. C. J. Godfray, C. Donnelly, G. Hewinson, M. Winter, and J. Wood, "Bovine tb strategy review," 2018.

[50] C. Humphries, "Latency: A sleeping giant," *Nature*, vol. 502, no. 7470, pp. S14–S15, 2013.

[51] P. E. Almeida Da Silva and J. C. Palomino, "Molecular basis and mechanisms of drug resistance in mycobacterium tuberculosis: classical and new drugs," *Journal of antimicrobial chemotherapy*, vol. 66, no. 7, pp. 1417–1430, 2011.

[52] C. Carranza, S. Pedraza-Sanchez, E. de Oyarzabal-Mendez, and M. Torres, "Diagnosis for latent tuberculosis infection: new alternatives," *Frontiers in immunology*, vol. 11, p. 2006, 2020.

[53] T. N. Jilani, A. Avula, Z. Gondal, and A. H. Siddiqui, "Active tuberculosis," 2018.

[54] L. M. Parsons, Ákos Somosközi, C. Gutierrez, E. Lee, C. N. Paramasivan, A. Abimiku, S. Spector, G. Roscigno, and J. Nkengasong, "Laboratory diagnosis of tuberculosis in resource-poor countries: Challenges and opportunities," *Clinical Microbiology Reviews*, vol. 24, no. 2, pp. 314–350, 2011.

[55] A. Koch, H. Cox, and V. Mizrahi, "Drug-resistant tuberculosis: challenges and opportunities for diagnosis and treatment," *Current Opinion in Pharmacology*, vol. 42, pp. 7–15, 2018.

[56] M. Kipiani, V. Mirtskhulava, N. Tukvadze, M. Magee, H. M. Blumberg, and R. R. Kempker, "Significant Clinical Impact of a Rapid Molecular Diagnostic Test (Genotype MTBDRplus Assay) to Detect Multidrug-Resistant Tuberculosis," *Clinical Infectious Diseases*, vol. 59, pp. 1559–1566, 08 2014.

[57] E. MacLean, M. Kohli, S. F. Weber, A. Suresh, S. G. Schumacher, C. M. Denkinger, and M. Pai, "Advances in molecular diagnosis of tuberculosis," *Journal of clinical microbiology*, vol. 58, no. 10, pp. 10–1128, 2020.

[58] B. Acharya, A. Acharya, S. Gautam, S. P. Ghimire, G. Mishra, N. Parajuli, and B. Sapkota, "Advances in diagnosis of tuberculosis: an update into molecular diagnosis of mycobacterium tuberculosis," *Molecular biology reports*, vol. 47, pp. 4065–4075, 2020.

[59] A. N. Unissa and L. E. Hanna, "Molecular mechanisms of action, resistance, detection to the first-line anti tuberculosis drugs: Rifampicin and pyrazinamide in the post whole genome sequencing era," *Tuberculosis*, vol. 105, pp. 96–107, 2017.

[60] J. I. Moliva, J. Turner, and J. B. Torrelles, "Immune responses to bacillus calmette–guerin vaccination: why do they fail to protect against mycobacterium tuberculosis?," *Frontiers in immunology*, vol. 8, p. 407, 2017.

[61] P. Nahid, S. E. Dorman, N. Alipanah, P. M. Barry, J. L. Brozek, A. Cattamanchi, L. H. Chaisson, R. E. Chaisson, C. L. Daley, M. Grzemska, *et al.*, "Official american thoracic society/centers for disease control and prevention/infectious diseases society of america clinical practice guidelines: treatment of drug-susceptible tuberculosis," *Clinical infectious diseases*, vol. 63, no. 7, pp. e147–e195, 2016.

[62] M. A. Espinal, A. Laszlo, L. Simonsen, F. Boulahbal, S. J. Kim, A. Reniero, S. Hoffner, H. L. Rieder, N. Binkin, C. Dye, *et al.*, "Global trends in resistance to antituberculosis drugs," *New England Journal of Medicine*, vol. 344, no. 17, pp. 1294–1303, 2001.

[63] F. Bardou, C. Raynaud, C. Ramos, M. A. Laneelle, and G. Laneelle, "Mechanism of isoniazid uptake in mycobacterium tuberculosis," *Microbiology*, vol. 144, no. 9, pp. 2539–2544, 1998.

[64] Y. Zhang, B. Heym, B. Allen, D. Young, and S. Cole, "The catalase—peroxidase gene and isoniazid resistance of mycobacterium tuberculosis," *Nature*, vol. 358, no. 6387, pp. 591–593, 1992.

[65] A. Banerjee, E. Dubnau, A. Quemard, V. Balasubramanian, K. S. Um, T. Wilson, D. Collins, G. De Lisle, and W. R. Jacobs Jr, "inha, a gene encoding a target for isoniazid and ethionamide in mycobacterium tuberculosis," *Science*, vol. 263, no. 5144, pp. 227–230, 1994.

[66] C. Vilcheze and W. R. Jacobs Jr, "Resistance to isoniazid and ethionamide in mycobacterium tuberculosis: genes, mutations, and causalities," *Molecular genetics of Mycobacteria*, pp. 431–453, 2014.

[67] M. H. Hazbon, M. Brimacombe, M. Bobadilla del Valle, M. Cavatore, M. I. Guerrero, M. Varma-Basil, H. Billman-Jacobe, C. Lavender, J. Fyfe, L. Garcia-Garcia, *et al.*, "Population genetics study of isoniazid resistance mutations and evolution of multidrug-resistant mycobacterium tuberculosis," *Antimicrobial agents and chemotherapy*, vol. 50, no. 8, pp. 2640–2649, 2006.

[68] G. S. Timmins and V. Deretic, "Mechanisms of action of isoniazid," *Molecular microbiology*, vol. 62, no. 5, pp. 1220–1227, 2006.

[69] S. Ramaswamy and J. M. Musser, "Molecular genetic basis of antimicrobial agent resistance inmycobacterium tuberculosis: 1998 update," *Tubercle and Lung disease*, vol. 79, no. 1, pp. 3–29, 1998.

[70] A. Telenti, P. Imboden, F. Marchesi, T. Schmidheini, and T. Bodmer, "Direct, automated detection of rifampin-resistant mycobacterium tuberculosis by polymerase chain reaction and single-strand conformation polymorphism analysis," *Antimicrobial agents and chemotherapy*, vol. 37, no. 10, pp. 2054–2058, 1993.

[71] B. P. Goldstein, "Resistance to rifampicin: a review," *The Journal of antibiotics*, vol. 67, no. 9, pp. 625–630, 2014.

[72] H. Safi, B. Sayers, M. H. Hazbon, and D. Alland, "Transfer of embb codon 306 mutations into clinical mycobacterium tuberculosis strains alters susceptibility to ethambutol, isoniazid, and rifampin," *Antimicrobial agents and chemotherapy*, vol. 52, no. 6, pp. 2027–2034, 2008.

[73] K. Takayama and J. O. Kilburn, "Inhibition of synthesis of arabinogalactan by ethambutol in mycobacterium smegmatis," *Antimicrobial agents and chemotherapy*, vol. 33, no. 9, pp. 1493–1499, 1989.

[74] K. Schubert, B. Sieger, F. Meyer, G. Giacomelli, K. Bohm, A. Rieblinger, L. Lindenthal, N. Sachs, G. Wanner, and M. Bramkamp, "The antituberculosis drug ethambutol selectively blocks apical growth in cmn group bacteria," *MBio*, vol. 8, no. 1, pp. 10–1128, 2017.

[75] X. Xiang, Z. Gong, W. Deng, Q. Sun, and J. Xie, "Mycobacterial ethambutol responsive genes and implications in antibiotics resistance," *Journal of Drug Targeting*, vol. 29, no. 3, pp. 284–293, 2021.

[76] S. Okamoto, A. Tamaru, C. Nakajima, K. Nishimura, Y. Tanaka, S. Tokuyama, Y. Suzuki, and K. Ochi, "Loss of a conserved 7-methylguanosine modification in 16s rrna confers low-level streptomycin resistance in bacteria," *Molecular microbiology*, vol. 63, no. 4, pp. 1096–1106, 2007.

[77] F. S. Spies, P. E. Almeida da Silva, M. O. Ribeiro, M. L. Rossetti, and A. Zaha, "Identification of mutations related to streptomycin resistance in clinical isolates of mycobacterium tuberculosis and possible involvement of efflux mechanism," *Antimicrobial agents and chemotherapy*, vol. 52, no. 8, pp. 2947–2949, 2008.

[78] N. Smittipat, T. Juthayothin, P. Billamas, S. Jaitrong, K. Rukseree, K. Dokladda, B. Chaiyasirinroje, A. Disratthakit, A. Chaiprasert, S. Mahasirimongkol, *et al.*, "Mutations in rrs, rpsl and gidb in streptomycin-resistant mycobacterium tuberculosis isolates from thailand," *Journal of global antimicrobial resistance*, vol. 4, pp. 5–10, 2016.

[79] A. S. Ginsburg, J. H. Grosset, and W. R. Bishai, "Fluoroquinolones, tuberculosis, and resistance," *The Lancet infectious diseases*, vol. 3, no. 7, pp. 432–442, 2003.

[80] S. Kabir, Z. Tahir, N. Mukhtar, M. Sohail, M. Saqalein, and A. Rehman, "Fluoroquinolone resistance and mutational profile of gyra in pulmonary mdr tuberculosis patients," *BMC pulmonary medicine*, vol. 20, pp. 1–6, 2020.

[81] L. Jugheli, N. Bzekalava, P. de Rijk, K. Fissette, F. Portaels, and L. Rigouts, "High level of cross-resistance between kanamycin, amikacin, and capreomycin among mycobacterium tuberculosis isolates from georgia and a close relation with mutations in the rrs gene," *Antimicrobial Agents and Chemotherapy*, vol. 53, no. 12, pp. 5064–5068, 2009.

[82] S. Yezli, *Molecular basis of biocide resistance and susceptibility in bacteria*. Cardiff University (United Kingdom), 2007.

[83] L. E. Via, S.-N. Cho, S. Hwang, H. Bang, S. K. Park, H. S. Kang, D. Jeon, S. Y. Min, T. Oh, Y. Kim, *et al.*, "Polymorphisms associated with resistance and cross-resistance to aminoglycosides and capreomycin in mycobacterium tuberculosis isolates from south korean patients with drug-resistant tuberculosis," *Journal of clinical microbiology*, vol. 48, no. 2, pp. 402–411, 2010.

[84] A. Engstrom, N. Perskvist, J. Werngren, S. E. Hoffner, and P. Jureen, "Comparison of clinical isolates and in vitro selected mutants reveals that tlya is not a sensitive genetic marker for capreomycin resistance in mycobacterium tuberculosis," *Journal of antimicrobial chemotherapy*, vol. 66, no. 6, pp. 1247–1254, 2011.

[85] M. A. Zaunbrecher, R. D. Sikes Jr, B. Metchock, T. M. Shinnick, and J. E. Posey, "Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in mycobacterium tuberculosis," *Proceedings of the National Academy of Sciences*, vol. 106, no. 47, pp. 20004–20009, 2009.

[86] A. E. DeBarber, K. Mdluli, M. Bosman, L.-G. Bekker, and C. E. Barry 3rd, "Ethionamide activation and sensitivity in multidrug-resistant mycobacterium tuberculosis," *Proceedings of the National Academy of Sciences*, vol. 97, no. 17, pp. 9677–9682, 2000.

[87] S. Thee, A. Garcia-Prats, P. Donald, A. Hesseling, and H. Schaaf, "A review of the use of ethionamide and prothionamide in childhood tuberculosis," *Tuberculosis*, vol. 97, pp. 126–136, 2016.

[88] K. Leung, C. Yip, Y. Yeung, K. Wong, W. Chan, M. Chan, and K. Kam, "Usefulness of resistant gene markers for predicting treatment outcome on second-line anti-tuberculosis drugs," *Journal of applied microbiology*, vol. 109, no. 6, pp. 2087–2094, 2010.

[89] J. Caminero, "Treatment of multidrug-resistant tuberculosis: evidence and controversies," *The International Journal of Tuberculosis and Lung Disease*, vol. 10, no. 8, pp. 829–837, 2006.

[90] Y. Li, F. Wang, L. Wu, M. Zhu, G. He, X. Chen, F. Sun, Q. Liu, X. Wang, and W. Zhang, "Cycloserine for treatment of multidrug-resistant tuberculosis: a retrospective cohort study in china," *Infection and drug resistance*, pp. 721–731, 2019.

[91] J. G. Jang and J. H. Chung, "Diagnosis and treatment of multidrug-resistant tuberculosis," *Yeungnam University Journal of Medicine*, vol. 37, no. 4, pp. 277–285, 2020.

[92] F. Coll, J. Phelan, G. A. Hill-Cawthorne, M. B. Nair, K. Mallard, S. Ali, A. M. Abdallah, S. Alghamdi, M. Alsomali, A. O. Ahmed, *et al.*, "Genome-wide analysis of multi-and extensively drug-resistant mycobacterium tuberculosis," *Nature genetics*, vol. 50, no. 2, pp. 307–316, 2018.

[93] F. Meacci, G. Orru, E. Iona, F. Giannoni, C. Piersimoni, G. Pozzi, L. Fattorini, and M. R. Oggioni, "Drug resistance evolution of a mycobacterium tuberculosis strain from a noncompliant patient," *Journal of clinical microbiology*, vol. 43, no. 7, pp. 3114–3120, 2005.

[94] A. L. Manson, K. A. Cohen, T. Abeel, C. A. Desjardins, D. T. Armstrong, C. E. Barry III, J. Brand, T. G. G. C. B. J. . J. P. . M. L. . N. D. . V. A. . C. G. H. . F. P. . H. D. . V. der Walt Martie 7 Hoffner Sven 13, S. B. Chapman, S.-N. Cho, *et al.*, "Genomic analysis of globally diverse mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance," *Nature genetics*, vol. 49, no. 3, pp. 395–402, 2017.

[95] G. Sun, T. Luo, C. Yang, X. Dong, J. Li, Y. Zhu, H. Zheng, W. Tian, S. Wang, C. E. Barry III, *et al.*, "Dynamic population changes in mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients," *The Journal of infectious diseases*, vol. 206, no. 11, pp. 1724–1733, 2012.

[96] C. B. Ford, R. R. Shah, M. K. Maeda, S. Gagneux, M. B. Murray, T. Cohen, J. C. Johnston, J. Gardy, M. Lipsitch, and S. M. Fortune, "Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis," *Nature genetics*, vol. 45, no. 7, pp. 784–790, 2013.

[97] S. H. Mariam, J. Werngren, J. Aronsson, S. Hoffner, and D. I. Andersson, "Dynamics of antibiotic resistant mycobacterium tuberculosis during long-term infection and antibiotic treatment," *PloS one*, vol. 6, no. 6, p. e21147, 2011.

[98] J. E. de Steenwinkel, M. T. ten Kate, G. J. de Knegt, K. Kremer, R. E. Aarnoutse, M. J. Boeree, H. A. Verbrugh, D. van Soolingen, and I. A. Bakker-Woudenberg, "Drug susceptibility of mycobacterium tuberculosis beijing genotype and association with mdr tb," *Emerging infectious diseases*, vol. 18, no. 4, p. 660, 2012.

[99] N. Dookie, S. Rambaran, N. Padayatchi, S. Mahomed, and K. Naidoo, "Evolution of drug resistance in mycobacterium tuberculosis: a review on the molecular determinants of resistance and implications for personalized care," *Journal of Antimicrobial Chemotherapy*, vol. 73, no. 5, pp. 1138–1151, 2018.

[100] T. G. Clark, K. Mallard, F. Coll, M. Preston, S. Assefa, D. Harris, S. Ogwang, F. Mumbowa, B. Kirenga, D. M. O'Sullivan, *et al.*, "Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing," *PloS one*, vol. 8, no. 12, p. e83012, 2013.

[101] Y. Zhang and W. Yew, "Mechanisms of drug resistance in mycobacterium tuberculosis: update 2015," *The International Journal of Tuberculosis and Lung Disease*, vol. 19, no. 11, pp. 1276–1289, 2015.

[102] S. H. Gillespie, "Evolution of drug resistance in mycobacterium tuberculosis: clinical and molecular perspective," *Antimicrobial agents and chemotherapy*, vol. 46, no. 2, pp. 267–274, 2002.

[103] T. M. Walker, M. Merker, A. M. Knoblauch, P. Helbling, O. D. Schoch, M. J. van der Werf, K. Kranzer, L. Fiebig, S. Kroger, W. Haas, *et al.*, "A cluster of multidrug-resistant mycobacterium tuberculosis among patients arriving in europe from the horn of africa: a molecular epidemiological study," *The Lancet Infectious Diseases*, vol. 18, no. 4, pp. 431–440, 2018.

[104] L. Fenner, M. Egger, T. Bodmer, E. Altpeter, M. Zwahlen, K. Jaton, G. E. Pfyffer, S. Borrell, O. Dubuis, T. Bruderer, *et al.*, "Effect of mutation and genetic background on drug resistance in mycobacterium tuberculosis," *Antimicrobial agents and chemotherapy*, vol. 56, no. 6, pp. 3047–3053, 2012.

[105] B. Motavaf, N. Keshavarz, F. Ghorbanian, S. Firuzabadi, F. Hosseini, and S. Z. Bostanabad, "Detection of genomic mutations in katg and rpob genes among multidrug-resistant mycobacterium tuberculosis isolates from tehran, iran," *New Microbes and New Infections*, vol. 41, p. 100879, 2021.

[106] O. Emmanuel and O. Okamgba, "Phenotypic detection of rifampicin and isoniazid resistance pattern of mycobacterium tuberculosis auramine positive isolates in mtata, south africa," *Microbiol Infect Dis*, vol. 6, no. 2, pp. 1–4, 2022.

[107] B. Klotoe, S. Kacimi, E. Costa-Conceicao, H. Gomes, R. Barcellos, S. Panaiotov, D. Haj Slimene, N. Sikhayeva, S. Sengstake, A. Schuitema, *et al.*, "Genomic characterization of mdr/xdr-tb in kazakhstan by a combination of high-throughput methods predominantly shows the ongoing transmission of l2/beijing 94–32 central asian/russian clusters," *BMC Infectious Diseases*, vol. 19, no. 1, pp. 1–12, 2019.

[108] S. Mak, Y. Xu, and J. R. Nodwell, "The expression of antibiotic resistance genes in antibiotic-producing bacteria," *Molecular microbiology*, vol. 93, no. 3, pp. 391–402, 2014.

[109] J. M. Coughlin, J. D. Rudolf, E. Wendt-Pienkowski, L. Wang, C. Unsin, U. Galm, D. Yang, M. Tao, and B. Shen, "Blmb and tlmb provide resistance to the bleomycin family of antitumor antibiotics by n-acetylating metal-free bleomycin, tallysomycin, phleomycin, and zorbamycin," *Biochemistry*, vol. 53, no. 44, pp. 6901–6909, 2014.

[110] W. Li, M. Sharma, and P. Kaur, "The drrab efflux system of streptomyces peucetius is a multidrug transporter of broad substrate specificity," *Journal of biological chemistry*, vol. 289, no. 18, pp. 12633–12646, 2014.

[111] E. E. Chufan, H.-M. Sim, and S. V. Ambudkar, "Molecular basis of the polyspecificity of p-glycoprotein (abcb1): recent biochemical and structural studies," *Advances in cancer research*, vol. 125, pp. 71–96, 2015.

[112] J. D. Rudolf, L. Bigelow, C. Chang, M. E. Cuff, J. R. Lohman, C.-Y. Chang, M. Ma, D. Yang, S. Clancy, G. Babnigg, *et al.*, "Crystal structure of the zorbamycin-binding protein zbma, the primary self-resistance element in streptomyces flavoviridis atcc21892," *Biochemistry*, vol. 54, no. 45, pp. 6842–6851, 2015.

[113] C. Laing, C. Buchanan, E. N. Taboada, Y. Zhang, A. Kropinski, A. Villegas, J. E. Thomas, and V. P. Gannon, "Pan-genome sequence analysis using panseq: an online tool for the rapid analysis of core and accessory genomic regions," *BMC bioinformatics*, vol. 11, pp. 1–14, 2010.

[114] F. Prija and R. Prasad, "Drrc protein of streptomyces peucetius removes daunorubicin from intercalated dnri promoter," *Microbiological Research*, vol. 202, pp. 30–35, 2017.

[115] F. Zakham, T. Sironen, O. Vapalahti, and R. Kant, "Pan and core genome analysis of 183 mycobacterium tuberculosis strains revealed a high inter-species diversity among the human adapted strains," *Antibiotics*, vol. 10, no. 5, p. 500, 2021.

[116] T. Yang, J. Zhong, J. Zhang, C. Li, X. Yu, J. Xiao, X. Jia, N. Ding, G. Ma, G. Wang, *et al.*, "Pan-genomic study of mycobacterium tuberculosis reflecting the primary/secondary genes, generality/individuality, and the interconversion through copy number variations," *Frontiers in microbiology*, vol. 9, p. 1886, 2018.

[117] A. M. Negrete-Paz, G. Vázquez-Marrufo, A. Gutiérrez-Moraga, and M. S. Vázquez-Garcidueñas, "Pangenome reconstruction of mycobacterium tuberculosis as a guide to reveal genomic features associated with strain clinical phenotype," *Microorganisms*, vol. 11, no. 6, p. 1495, 2023.

[118] A. Tiwari *et al.*, "Applications of bioinformatics tools to combat the antibiotic resistance," in *2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI)*, pp. 96–98, IEEE, 2015.

[119] S. K. David, A. T. Saeb, M. Rafiullah, and K. Rubeaan, "Classification techniques and data mining tools used in medical bioinformatics," in *Big data governance and perspectives in knowledge management*, pp. 105–126, IGI Global, 2019.

[120] T. A. Kumbhare and S. V. Chobe, "An overview of association rule mining algorithms," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 1, pp. 927–930, 2014.

[121] S. L. Patil, "Survey of data mining techniques in healthcare," *International Research Journal of Innovative Engineering, Volume1*, no. 9, pp. 1–3, 2015.

[122] P. R. Harper, "A review and comparison of classification algorithms for medical decision making," *Health policy*, vol. 71, no. 3, pp. 315–331, 2005.

[123] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, "Data mining in healthcare and biomedicine: a survey of the literature," *Journal of medical systems*, vol. 36, pp. 2431–2448, 2012.

[124] P. Sinha, M. Churpek, and C. Calfee, "Machine learning classifier models can identify ards phenotypes using readily available clinical data," in *A15. CRITICAL CARE: BRAVE NEW WORLD-NEW INSIGHTS FROM CLINICAL TRIALS AND OBSERVATIONAL COHORTS*, pp. A1014–A1014, American Thoracic Society, 2019.

[125] K. Sharmila and S. Vethamanickam, "Survey on data mining algorithm and its application in healthcare sector using hadoop platform," *International Journal of Emerging Technology and Advanced Engineering*, vol. 5, no. 1, pp. 567–571, 2015.

[126] J. J. Tapia, E. Morett, and E. E. Vallejo, "A clustering genetic algorithm for genomic data mining," in *Foundations of Computational Intelligence Volume 4: Bio-Inspired Data Mining*, pp. 249–275, Springer, 2009.

[127] N. Kausar, A. Abdullah, B. B. Samir, S. Palaniappan, B. S. AlGhamdi, and N. Dey, "Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease," *Journal of Medical Imaging and Health Informatics*, vol. 6, no. 1, pp. 78–87, 2016.

[128] C. Zhang and S. Zhang, *Association rule mining: models and algorithms*. Springer, 2002.

[129] M. Babaie, S. Kalra, A. Sriram, C. Mitcheltree, S. Zhu, A. Khatami, S. Rahnamayan, and H. R. Tizhoosh, "Classification and retrieval of digital pathology scans: A new dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 8–16, 2017.

[130] I. E. Emre, N. Erol, Y. I. Ayhan, Y. Ozkan, and c. Erol, "The analysis of the effects of acute rheumatic fever in childhood on cardiac disease with data mining," *International journal of medical informatics*, vol. 123, pp. 68–75, 2019.

[131] Y. Xing, J. Wang, Z. Zhao, *et al.*, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," in *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, pp. 868–872, IEEE, 2007.

[132] T.-H. Cheng, C.-P. Wei, and V. S. Tseng, "Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches," in *19th IEEE symposium on computer-based medical systems (CBMS'06)*, pp. 165–170, IEEE, 2006.

[133] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics and Informatics*, vol. 36, pp. 82–93, 2019.

[134] G. Santos-Garcia, G. Varela, N. Novoa, and M. F. Jimenez, "Prediction of postoperative morbidity after lung resection using an artificial neural network ensemble," *Artificial intelligence in medicine*, vol. 30, no. 1, pp. 61–69, 2004.

[135] P.-C. Hsieh, C.-F. Cheng, C.-W. Wu, I. Tzeng, C.-Y. Kuo, P.-S. Hsu, C.-T. Lee, M.-C. Yu, C.-C. Lan, *et al.*, "Combination of acupoints in treating patients with chronic obstructive pulmonary disease: an apriori algorithm-based association rule analysis," *Evidence-Based Complementary and Alternative Medicine*, vol. 2020, 2020.

[136] D.-I. Curiac, G. Vasile, O. Banias, C. Volosencu, and A. Albu, "Bayesian network model for diagnosis of psychiatric diseases," in *Proceedings of the ITI 2009 31st international conference on information technology interfaces*, pp. 61–66, IEEE, 2009.

[137] S. G. Alonso, I. de La Torre-Diez, S. Hamrioui, M. Lopez-Coronado, D. C. Barreno, L. M. Nozaleda, and M. Franco, "Data mining algorithms and techniques in mental health: a systematic review," *Journal of medical systems*, vol. 42, pp. 1–15, 2018.

[138] T. A. Gameel, S. Rady, and S. M. Kamal, "Risks and predictors of non-alcoholic liver disease progression using association rules mining.," *Int. J. Online Biomed. Eng.*, vol. 16, no. 6, pp. 61–71, 2020.

[139] C.-L. Chang and C.-H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis," *Expert Systems with Applications*, vol. 36, no. 2, pp. 4035–4041, 2009.

[140] A. K. Verma, S. Pal, and S. Kumar, "Classification of skin disease using ensemble data mining techniques," *Asian Pacific journal of cancer prevention: APJCP*, vol. 20, no. 6, p. 1887, 2019.

[141] N. Rane and M. Rao, "Association rule mining on type 2 diabetes using fp-growth association rule," *International journal of engineering and computer science*, vol. 2, no. 8, p. 4, 2013.

[142] P. M. Shakeel, S. Baskar, V. S. Dhulipala, and M. M. Jaber, "Cloud based framework for diagnosis of diabetes mellitus using k-means clustering," *Health information science and systems*, vol. 6, pp. 1–7, 2018.

[143] S. Z. Hassan and B. Verma, "A hybrid data mining approach for knowledge extraction and classification in medical databases," in *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*, pp. 503–510, IEEE, 2007.

[144] D. Umesh and B. Ramachandra, "Association rule mining based predicting breast cancer recurrence on seer breast cancer data," in *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, pp. 376–380, IEEE, 2015.

[145] Z. Momeni, E. Hassanzadeh, M. S. Abadeh, and R. Bellazzi, "A survey on single and multi omics data mining methods in cancer data classification," *Journal of Biomedical Informatics*, vol. 107, p. 103466, 2020.

[146] A. K. Srivastava, K. Jeberson, and W. Jeberson, "A systematic review on data mining application in parkinson's disease," *Neuroscience Informatics*, vol. 2, no. 4, p. 100064, 2022.

[147] N. Nahar, F. Ara, M. A. I. Neloy, A. Biswas, M. S. Hossain, and K. Andersson, "Feature selection based machine learning to improve prediction of parkinson disease," in *Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings 14*, pp. 496–508, Springer, 2021.

[148] E. G. Giannopoulou, V. P. Kemerlis, M. Polemis, J. Papaparaskevas, A. C. Vatopoulos, and M. Vazirgiannis, "A large scale data mining approach to antibiotic resistance surveillance," in *Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)*, pp. 439–444, IEEE, 2007.

[149] S. Mutalib, N. A. Ali, S. A. Rahman, and A. Mohamed, "An exploratory study in classification methods for patients' dataset," in *2009 2nd Conference on Data Mining and Optimization*, pp. 79–83, IEEE, 2009.

[150] M. Gerontini, M. Vazirgiannis, A. C. Vatopoulos, and M. Polemis, "Predictions in antibiotics resistance and nosocomial infections monitoring," in *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 1–6, IEEE, 2011.

[151] K. Vougas, T. Sakellaropoulos, A. Kotsinas, G.-R. P. Foukas, A. Ntargaras, F. Koinis, A. Polyzos, V. Myrianthopoulos, H. Zhou, S. Narang, *et al.*, "Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining," *Pharmacology, therapeutics*, vol. 203, p. 107395, 2019.

[152] S. Kouchaki, Y. Yang, T. M. Walker, A. Sarah Walker, D. J. Wilson, T. E. Peto, D. W. Crook, C. Consortium, and D. A. Clifton, "Application of machine learning techniques to tuberculosis drug resistance analysis," *Bioinformatics*, vol. 35, no. 13, pp. 2276–2282, 2019.

[153] T. Uccar and A. Karahoca, "Predicting existence of mycobacterium tuberculosis on patients using data mining approaches," *Procedia Computer Science*, vol. 3, pp. 1404–1411, 2011.

[154] D. Ashwini and S. Seema, "Machine learning approach to detect tuberculosis in patients with or without hiv co-infection-a survey," 2015.

[155] S. Sathitratanacheewin, P. Sunanta, and K. Pongpirul, "Deep learning for automated classification of tuberculosis-related chest x-ray: dataset distribution shift limits diagnostic performance generalizability," *Heliyon*, vol. 6, no. 8, 2020.

[156] S. Kulshrestha, S. Panda, D. Nayar, V. Dohe, and A. Jarali, "Prediction of antimicrobial resistance for disease-causing agents using machine learning," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 972–975, IEEE, 2018.

[157] C. A. Bobak, A. J. Titus, and J. E. Hill, "Comparison of common machine learning models for classification of tuberculosis using transcriptional biomarkers from integrated datasets," *Applied Soft Computing*, vol. 74, pp. 264–273, 2019.

[158] K. C. Ng, J. C. S. Ngabonziza, P. Lempens, B. C. de Jong, F. van Leth, and C. J. Meehan, "Bridging the tb data gap: in silico extraction of rifampicin-resistant tuberculosis diagnostic test results from whole genome sequence data," *PeerJ*, vol. 7, p. e7564, 2019.

[159] D. Aytan-Aktug, P. T. L. C. Clausen, V. Bortolaia, F. M. Aarestrup, and O. Lund, "Prediction of acquired antimicrobial resistance for multiple bacterial species using neural networks," *Msystems*, vol. 5, no. 1, pp. 10–1128, 2020.

[160] S. Jamal, M. Khubaib, R. Gangwar, S. Grover, A. Grover, and S. E. Hasnain, "Artificial intelligence and machine learning based prediction of resistant and susceptible mutations in mycobacterium tuberculosis," *Scientific reports*, vol. 10, no. 1, p. 5487, 2020.

[161] M. Harahap, A. Husein, S. Aisyah, F. Lubis, and B. Wijaya, "Mining association rule based on the diseases population for recommendation of medicine need," in *Journal of Physics: Conference Series*, vol. 1007, p. 012017, IOP Publishing, 2018.

[162] M. Tandan, Y. Acharya, S. Pokharel, and M. Timilsina, "Discovering symptom patterns of covid-19 patients using association rule mining," *Computers in biology and medicine*, vol. 131, p. 104249, 2021.

[163] R. Veroneze, S. Cruz Tfaile Corbi, B. Roque da Silva, C. de S. Rocha, C. V. Maurer-Morelli, S. R. Perez Orrico, J. A. Cirelli, F. J. Von Zuben, and R. Mantuaneli Scarel-Caminaga, "Using association rule mining to jointly detect clinical features and differentially expressed genes related to chronic inflammatory diseases," *Plos one*, vol. 15, no. 10, p. e0240269, 2020.

[164] V. C. D. Micheletti, A. L. Kritski, and J. U. Braga, "Clinical features and treatment outcomes of patients with drug-resistant and drug-sensitive tuberculosis: a historical cohort study in porto alegre, brazil," *PLoS One*, vol. 11, no. 8, p. e0160109, 2016.

[165] B. Baya, C. J. Achenbach, B. Kone, Y. Toloba, D. K. Dabitao, B. Diarra, D. Goita, S. Diabate, M. Maiga, D. Soumare, *et al.*, "Clinical risk factors associated with multidrug-resistant tuberculosis (mdr-tb) in mali," *International Journal of Infectious Diseases*, vol. 81, pp. 149–155, 2019.

[166] S. Ali, M. T. Khan, A. S. Khan, N. Mohammad, M. M. Khan, S. Ahmad, S. Noor, A. Jabbar, C. Daire, and F. Hassan, "Prevalence of multi-drug resistant mycobacterium tuberculosis in khyber pakhtunkhwa–a high tuberculosis endemic area of pakistan," *Polish journal of microbiology*, vol. 69, no. 2, pp. 133–137, 2020.

[167] J. A. Tornheim, E. Intini, A. Gupta, and Z. F. Udwadia, "Clinical features associated with linezolid resistance among multidrug resistant tuberculosis patients at a tertiary care hospital in mumbai, india," *Journal of clinical tuberculosis and other mycobacterial diseases*, vol. 20, p. 100175, 2020.

[168] T. Takii, K. Seki, Y. Wakabayashi, Y. Morishige, T. Sekizuka, A. Yamashita, K. Kato, K. Uchimura, A. Ohkado, N. Keicho, *et al.*, "Whole-genome sequencing-based epidemiological analysis of anti-tuberculosis drug resistance genes in japan in 2007: application of the genome research for asian tuberculosis (great) database," *Scientific Reports*, vol. 9, no. 1, p. 12823, 2019.

[169] J. R. Coelho, J. A. Carricco, D. Knight, J.-L. Martinez, I. Morrissey, M. R. Oggioni, and A. T. Freitas, "The use of machine learning methodologies to analyse antibiotic and biocide susceptibility in staphylococcus aureus," *PLoS One*, vol. 8, no. 2, p. e55582, 2013.

[170] J. Shi, R. Su, D. Zheng, Y. Zhu, X. Ma, S. Wang, H. Li, and D. Sun, "Pyrazinamide resistance and mutation patterns among multidrug-resistant mycobacterium tuberculosis from henan province," *Infection and drug resistance*, pp. 2929–2941, 2020.

[171] X. Chen, G. He, S. Wang, S. Lin, J. Chen, and W. Zhang, "Evaluation of whole-genome sequence method to diagnose resistance of 13 anti-tuberculosis drugs and characterize resistance genes in clinical multi-drug resistance mycobacterium tuberculosis isolates from china," *Frontiers in microbiology*, vol. 10, p. 1741, 2019.

[172] A. Sowajassatakul, T. Prammananan, A. Chaiprasert, and S. Phunpruch, "Molecular characterization of amikacin, kanamycin and capreomycin resistance in m/xdr-tb strains isolated in thailand," *BMC microbiology*, vol. 14, pp. 1–7, 2014.

[173] S. B. Georghiou, M. Magana, R. S. Garfein, D. G. Catanzaro, A. Catanzaro, and T. C. Rodwell, "Evaluation of genetic mutations associated with

mycobacterium tuberculosis resistance to amikacin, kanamycin and capreomycin: a systematic review," *PloS one*, vol. 7, no. 3, p. e33275, 2012.

[174] X. Yin and Z. Yu, "Mutation characterization of gyra and gyrb genes in levofloxacin-resistant mycobacterium tuberculosis clinical isolates from guangdong province in china," *Journal of Infection*, vol. 61, no. 2, pp. 150–154, 2010.

[175] Y. Pang, Z. Zhang, Y. Wang, S. Wang, Y. Song, B. Zhao, Y. Zhou, X. Ou, Q. Li, H. Xia, *et al.*, "Genotyping and prevalence of pyrazinamide-and moxifloxacin-resistant tuberculosis in china, 2000 to 2010," *Antimicrobial agents and chemotherapy*, vol. 61, no. 2, pp. 10–1128, 2017.

[176] J. Chen, S. Zhang, P. Cui, W. Shi, W. Zhang, and Y. Zhang, "Identification of novel mutations associated with cycloserine resistance in mycobacterium tuberculosis," *Journal of Antimicrobial Chemotherapy*, vol. 72, no. 12, pp. 3272–3276, 2017.

[177] P. Beckert, D. Hillemann, T. A. Kohl, J. Kalinowski, E. Richter, S. Niemann, and S. Feuerriegel, "rplc t460c identified as a dominant mutation in linezolid-resistant mycobacterium tuberculosis strains," *Antimicrobial agents and chemotherapy*, vol. 56, no. 5, pp. 2743–2745, 2012.

[178] F. Maruri, T. R. Sterling, A. W. Kaiga, A. Blackman, Y. F. van der Heijden, C. Mayer, E. Cambau, and A. Aubry, "A systematic review of gyrase mutations associated with fluoroquinolone-resistant mycobacterium tuberculosis and a proposed gyrase numbering system," *Journal of Antimicrobial Chemotherapy*, vol. 67, no. 4, pp. 819–831, 2012.

[179] A. Z. Reeves, P. J. Campbell, R. Sultana, S. Malik, M. Murray, B. B. Plikaytis, T. M. Shinnick, and J. E. Posey, "Aminoglycoside cross-resistance in mycobacterium tuberculosis due to mutations in the 5 untranslated region of whib7," *Antimicrobial agents and chemotherapy*, vol. 57, no. 4, pp. 1857–1865, 2013.

[180] B. Panic, J. Klemenc, and M. Nagode, "Optimizing the estimation of a histogram-bin width—application to the multivariate mixture-model estimation," *Mathematics*, vol. 8, no. 7, p. 1090, 2020.

[181] N. N. Xiong, Y. Shen, K. Yang, C. Lee, and C. Wu, "Color sensors and their applications based on real-time color image segmentation for cyber physical systems," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, pp. 1–16, 2018.

[182] W. Mulu, D. Mekkonnen, M. Yimer, A. Admassu, and B. Abera, "Risk factors for multidrug resistant tuberculosis patients in amhara national regional state.," *African health sciences*, vol. 15, no. 2, pp. 368–377, 2015.

[183] K. Mehari, T. Asmelash, H. Hailekiros, T. Wubayehu, H. Godefay, T. Araya, and M. Saravanan, "Prevalence and factors associated with multidrug-resistant tuberculosis (mdr-tb) among presumptive mdr-tb patients in tigray region, northern ethiopia," *Canadian Journal of Infectious Diseases and Medical Microbiology*, vol. 2019, 2019.

[184] F. Mekonnen, B. Tessema, F. Moges, A. Gelaw, S. Eshetie, and G. Kumera, "Multidrug resistant tuberculosis: prevalence and risk factors in districts of metema and west armachiho, northwest ethiopia," *BMC infectious diseases*, vol. 15, pp. 1–6, 2015.

[185] M. Rifat, A. H. Milton, J. Hall, C. Oldmeadow, M. A. Islam, A. Husain, M. W. Akhanda, and B. N. Siddiquea, "Development of multidrug resistant tuberculosis in bangladesh: a case-control study on risk factors," *PloS one*, vol. 9, no. 8, p. e105214, 2014.

[186] N. Lomtadze, R. Aspindzelashvili, M. Janjgava, V. Mirtskhulava, A. Wright, H. Blumberg, and A. Salakaia, "Prevalence and risk factors for multidrug-resistant tuberculosis in the republic of georgia: a population-based study," *The International journal of tuberculosis and lung disease*, vol. 13, no. 1, pp. 68–73, 2009.

[187] C. F. McQuaid, K. C. Horton, A. S. Dean, G. M. Knight, and R. G. White, "The risk of multidrug-or rifampicin-resistance in males versus females with tuberculosis," *European Respiratory Journal*, vol. 56, no. 3, 2020.

[188] G. S. John, N. Brot, J. Ruan, H. Erdjument-Bromage, P. Tempst, H. Weissbach, and C. Nathan, "Peptide methionine sulfoxide reductase from escherichia coli and mycobacterium tuberculosis protects bacteria against oxidative damage from reactive nitrogen intermediates," *Proceedings of the National Academy of Sciences*, vol. 98, no. 17, pp. 9901–9906, 2001.

[189] W. L. Lee, B. Gold, C. Darby, N. Brot, X. Jiang, L. P. S. De Carvalho, D. Wellner, G. St. John, W. R. Jacobs Jr, and C. Nathan, "Mycobacterium tuberculosis expresses methionine sulphoxide reductases a and b that protect from killing by nitrite and hypochlorite," *Molecular microbiology*, vol. 71, no. 3, pp. 583–593, 2009.

[190] A. Gurvitz, J. K. Hiltunen, and A. J. Kastaniotis, "Function of heterologous mycobacterium tuberculosis inha, a type 2 fatty acid synthase enzyme involved in extending c20 fatty acids to c60-to-c90 mycolic acids, during de novo lipoic acid synthesis in saccharomyces cerevisiae," *Applied and Environmental Microbiology*, vol. 74, no. 16, pp. 5078–5085, 2008.

[191] Y. H. Lee, K. H. Nam, and J. D. Helmann, "A mutation of the rna polymerase b subunit (rpoc) confers cephalosporin resistance in bacillus subtilis," *Antimicrobial agents and chemotherapy*, vol. 57, no. 1, pp. 56–65, 2013.

[192] D. J. Farrell, M. Castanheira, and I. Chopra, "Characterization of global patterns and the genetics of fusidic acid resistance," *Clinical infectious diseases*, vol. 52, no. 7, pp. S487–S492, 2011.

[193] T. Wassenaar, D. Ussery, L. Nielsen, and H. Ingmer, "Review and phylogenetic analysis of qac genes that reduce susceptibility to quaternary ammonium compounds in staphylococcus species," *European Journal of Microbiology and Immunology*, vol. 5, no. 1, pp. 44–61, 2015.

[194] C. Achilli, A. Ciana, and G. Minetti, "The discovery of methionine sulfoxide reductase enzymes: An historical account and future perspectives," *Biofactors*, vol. 41, no. 3, pp. 135–152, 2015.

[195] J. L. Kandler, A. D. Mercante, T. L. Dalton, M. N. Ezewudo, L. S. Cowan, S. P. Burns, B. Metchock, P. Cegielski, and J. E. Posey, "Validation of novel mycobacterium tuberculosis isoniazid resistance mutations not detectable by common molecular tests," *Antimicrobial agents and chemotherapy*, vol. 62, no. 10, pp. 10–1128, 2018.

[196] R. Wallace and D. Griffith, "Antimycrobial agents in kasper dl, braunwald e (eds), harrison's principles of internal medicine," 2004.

[197] Q.-j. Li, W.-w. Jiao, Q.-q. Yin, F. Xu, J.-q. Li, L. Sun, J. Xiao, Y.-j. Li, I. Mokrousov, H.-r. Huang, *et al.*, "Compensatory mutations of rifampin resistance are associated with transmission of multidrug-resistant mycobacterium tuberculosis beijing genotype strains in china," *Antimicrobial agents and chemotherapy*, vol. 60, no. 5, pp. 2807–2812, 2016.

[198] C. Cicek-Saydam, C. Cavusoglu, D. Burhanoglu, S. Hilmioglu, N. Ozkalay, and A. Bilgic, "In vitro susceptibility of mycobacterium tuberculosis to fusidic acid," *Clinical Microbiology and Infection*, vol. 7, no. 12, pp. 700–702, 2001.

[199] W.-L. Huang, T.-L. Chi, M.-H. Wu, and R. Jou, "Performance assessment of the genotype mtbdr sl test and dna sequencing for detection of second-line and ethambutol drug resistance among patients infected with multidrug-resistant mycobacterium tuberculosis," *Journal of clinical microbiology*, vol. 49, no. 7, pp. 2502–2508, 2011.

[200] S. S. Hegde, M. W. Vetting, S. L. Roderick, L. A. Mitchenall, A. Maxwell, H. E. Takiff, and J. S. Blanchard, "A fluoroquinolone resistance protein from mycobacterium tuberculosis that mimics dna," *Science*, vol. 308, no. 5727, pp. 1480–1483, 2005.

[201] S. K. Sharma and A. Mohan, "Multidrug-resistant tuberculosis: a menace that threatens to destabilize tuberculosis control," *Chest*, vol. 130, no. 1, pp. 261–272, 2006.

[202] V. E. Escuyer, M.-A. Lety, J. B. Torrelles, K.-H. Khoo, J.-B. Tang, C. D. Rithner, C. Frehel, M. R. McNeil, P. J. Brennan, and D. Chatterjee, "The role of the emba and embb gene products in the biosynthesis of the terminal hexaarabinofuranosyl motif of mycobacterium smegmatisarabinogalactan," *Journal of Biological Chemistry*, vol. 276, no. 52, pp. 48854–48862, 2001.

[203] S. Berg, D. Kaur, M. Jackson, and P. J. Brennan, "The glycosyltransferases of mycobacterium tuberculosis—roles in the synthesis of arabinogalactan, lipoarabinomannan, and other glycoconjugates," *Glycobiology*, vol. 17, no. 6, pp. 35R–56R, 2007.

[204] S. Chakraborty, T. Gruber, C. E. Barry III, H. I. Boshoff, and K. Y. Rhee, "Para-aminosalicylic acid acts as an alternative substrate of folate metabolism in mycobacterium tuberculosis," *Science*, vol. 339, no. 6115, pp. 88–91, 2013.

[205] L. Kremer, L. G. Dover, S. Carrere, K. M. Nampoothiri, S. Lesjean, A. K. Brown, P. J. Brennan, D. E. Minnikin, C. Locht, and G. S. Besra, "Mycolic acid biosynthesis and enzymic characterization of the β-ketoacyl-acp synthase a-condensing enzyme from mycobacterium tuberculosis," *Biochemical Journal*, vol. 364, no. 2, pp. 423–430, 2002.

[206] J. A. Ainsa, E. Perez, V. Pelicic, F.-X. Berthet, B. Gicquel, and C. Martin, "Aminoglycoside 2-n-acetyltransferase genes are universally present in mycobacteria: characterization of the aac (2)-ic gene from mycobacterium tuberculosis and the aac (2)-id gene from mycobacterium smegmatis," *Molecular microbiology*, vol. 24, no. 2, pp. 431–441, 1997.

[207] G. L. Newton, T. Koledin, B. Gorovitz, M. Rawat, R. C. Fahey, and Y. Av-Gay, "The glycosyltransferase gene encoding the enzyme catalyzing the first

step of mycothiol biosynthesis (msha)," *Journal of Bacteriology*, vol. 185, no. 11, pp. 3476–3479, 2003.

[208] R. Colangeli, D. Helb, S. Sridharan, J. Sun, M. Varma-Basil, M. H. Hazbon, R. Harbacheuski, N. J. Megjugorac, W. R. Jacobs Jr, A. Holzenburg, *et al.*, "The mycobacterium tuberculosis inia gene is essential for activity of an efflux pump that confers drug tolerance to both isoniazid and ethambutol," *Molecular microbiology*, vol. 55, no. 6, pp. 1829–1840, 2005.

[209] M. A. Sorensen, J. Fricke, and S. Pedersen, "Ribosomal protein s1 is required for translation of most, if not all, natural mrnas in escherichia coli in vivo," *Journal of molecular biology*, vol. 280, no. 4, pp. 561–569, 1998.

[210] A. Aubry, X.-S. Pan, L. M. Fisher, V. Jarlier, and E. Cambau, "Mycobacterium tuberculosis dna gyrase: interaction with quinolones and correlation with antimycobacterial drug activity," *Antimicrobial agents and chemotherapy*, vol. 48, no. 4, pp. 1281–1288, 2004.

[211] D. M. Mikheil, D. C. Shippy, N. M. Eakley, O. E. Okwumabua, and A. A. Fadl, "Deletion of gene encoding methyltransferase (gidb) confers high-level antimicrobial resistance in salmonella," *The Journal of antibiotics*, vol. 65, no. 4, pp. 185–192, 2012.

[212] S. Zeng, K. Soetaert, F. Ravon, M. Vandeput, D. Bald, J.-M. Kauffmann, V. Mathys, R. Wattiez, and V. Fontaine, "Isoniazid bactericidal activity involves electron transport chain perturbation," *Antimicrobial agents and chemotherapy*, vol. 63, no. 3, pp. 10–1128, 2019.

[213] H. Zhang, J.-Y. Deng, L.-J. Bi, Y.-F. Zhou, Z.-P. Zhang, C.-G. Zhang, Y. Zhang, and X.-E. Zhang, "Characterization of mycobacterium tuberculosis nicotinamidase/pyrazinamidase," *The FEBS journal*, vol. 275, no. 4, pp. 753–762, 2008.

[214] M. W. Fraaije, N. M. Kamerbeek, A. J. Heidekamp, R. Fortin, and D. B. Janssen, "The prodrug activator etaa from mycobacterium tuberculosis is a

baeyer-villiger monooxygenase," *Journal of Biological Chemistry*, vol. 279, no. 5, pp. 3354–3360, 2004.

[215] H. Jia, H. Chu, G. Dai, T. Cao, and Z. Sun, "Rv1258c acts as a drug efflux pump and growth controlling factor in mycobacterium tuberculosis," *Tuberculosis*, vol. 133, p. 102172, 2022.

[216] S. Kumar, A. Parvathi, R. L. Hernandez, K. M. Cadle, and M. F. Varela, "Identification of a novel udp-n-acetylglucosamine enolpyruvyl transferase (mura) from vibrio fischeri that confers high fosfomycin resistance in escherichia coli," *Archives of microbiology*, vol. 191, pp. 425–429, 2009.

[217] D. Alland, I. Kramnik, T. R. Weisbrod, L. Otsubo, R. Cerny, L. P. Miller, W. R. Jacobs Jr, and B. R. Bloom, "Identification of differentially expressed mrna in prokaryotic organisms by customized amplification libraries (decal): the effect of isoniazid on gene expression in mycobacterium tuberculosis," *Proceedings of the National Academy of Sciences*, vol. 95, no. 22, pp. 13227–13232, 1998.

[218] F. Fassbinder, M. Kist, and S. Bereswill, "Structural and functional analysis of the riboflavin synthesis genes encoding gtp cyclohydrolase ii (riba), dhbp synthase (ribba), riboflavin synthase (ribc), and riboflavin deaminase/reductase (ribd) from helicobacter pylori strain p1," *FEMS microbiology letters*, vol. 191, no. 2, pp. 191–197, 2000.

[219] K. Madhusudan and V. Nagaraja, "Mycobacterium smegmatis dna gyrase: cloning and overexpression in escherichia coli," *Microbiology*, vol. 141, no. 12, pp. 3029–3037, 1995.

[220] M. A. Rahman, P. Sobia, V. P. Dwivedi, A. Bhawsar, D. K. Singh, P. Sharma, P. Moodley, L. Van Kaer, W. R. Bishai, and G. Das, "Mycobacterium tuberculosis tlya protein negatively regulates t helper (th) 1 and th17 differentiation and promotes tuberculosis pathogenesis," *Journal of Biological Chemistry*, vol. 290, no. 23, pp. 14407–14417, 2015.

[221] J. L. Doran, Y. Pang, K. E. Mdluli, A. J. Moran, T. C. Victor, R. W. Stokes, E. Mahenthiralingam, B. N. Kreiswirth, J. L. Butt, G. S. Baron, *et al.*, "Mycobacterium tuberculosis efpa encodes an efflux protein of the qaca transporter family," *Clinical Diagnostic Laboratory Immunology*, vol. 4, no. 1, pp. 23–32, 1997.

[222] L. E. Holberger and C. S. Hayes, "Ribosomal protein s12 and aminoglycoside antibiotics modulate a-site mrna cleavage and transfer-messenger rna activity in escherichia coli," *Journal of Biological Chemistry*, vol. 284, no. 46, pp. 32188–32200, 2009.

[223] R. Singh, B. Wiseman, T. Deemagarn, V. Jha, J. Switala, and P. C. Loewen, "Comparative study of catalase-peroxidases (katgs)," *Archives of biochemistry and biophysics*, vol. 471, no. 2, pp. 207–214, 2008.

[224] Y. N. Zhou, L. Lubkowska, M. Hui, S. Chen, J. Strathern, D. J. Jin, M. Kashlev, *et al.*, "Isolation and characterization of rna polymerase rpob mutations that alter transcription slippage during elongation in escherichia coli," *Journal of Biological Chemistry*, vol. 288, no. 4, pp. 2700–2710, 2013.

[225] A. R. Flores, L. M. Parsons, and M. S. Pavelka Jr, "Genetic analysis of the b-lactamases of mycobacterium tuberculosis and mycobacterium smegmatis and susceptibility to b-lactam antibiotics," *Microbiology*, vol. 151, no. 2, pp. 521–532, 2005.

[226] V. Mathys, R. Wintjens, P. Lefevre, J. Bertout, A. Singhal, M. Kiass, N. Kurepina, X.-M. Wang, B. Mathema, A. Baulard, *et al.*, "Molecular genetics of para-aminosalicylic acid resistance in clinical isolates and spontaneous mutants of mycobacterium tuberculosis," *Antimicrobial agents and chemotherapy*, vol. 53, no. 5, pp. 2100–2109, 2009.

[227] Y. Xu, H. Jia, H. Huang, Z. Sun, Z. Zhang, *et al.*, "Mutations found in embcab, embr, and ubia genes of ethambutol-sensitive and-resistant mycobacterium tuberculosis clinical isolates from china," *BioMed research international*, vol. 2015, 2015.

[228] M. H. Larsen, C. Vilcheze, L. Kremer, G. S. Besra, L. Parsons, M. Salfinger, L. Heifets, M. H. Hazbon, D. Alland, J. C. Sacchettini, *et al.*, "Overexpression of inha, but not kasa, confers resistance to isoniazid and ethionamide in mycobacterium smegmatis, m. bovis bcg and m. tuberculosis," *Molecular microbiology*, vol. 46, no. 2, pp. 453–466, 2002.

[229] R. S. Joshi, M. D. Jamdhade, M. S. Sonawane, and A. P. Giri, "Resistome analysis of mycobacterium tuberculosis: Identification of aminoglycoside 2'-nacetyltransferase (aac) as co-target for drug desigining," *Bioinformation*, vol. 9, no. 4, p. 174, 2013.

[230] T. Jagielski, Z. Bakula, K. Roeske, M. Kaminski, A. Napiorkowska, E. Augustynowicz-Kopec, Z. Zwolska, and J. Bielecki, "Detection of mutations associated with isoniazid resistance in multidrug-resistant mycobacterium tuberculosis clinical isolates," *Journal of Antimicrobial Chemotherapy*, vol. 69, no. 9, pp. 2369–2375, 2014.

[231] M. Zhang, J. Yue, Y.-p. Yang, H.-m. Zhang, J.-q. Lei, R.-l. Jin, X.-l. Zhang, and H.-h. Wang, "Detection of mutations associated with isoniazid resistance in mycobacterium tuberculosis isolates from china," *Journal of clinical microbiology*, vol. 43, no. 11, pp. 5477–5482, 2005.

[232] Y. Tan, Z. Hu, T. Zhang, X. Cai, H. Kuang, Y. Liu, J. Chen, F. Yang, K. Zhang, S. Tan, *et al.*, "Role of pnca and rpsa gene sequencing in detection of pyrazinamide resistance in mycobacterium tuberculosis isolates from southern china," *Journal of clinical microbiology*, vol. 52, no. 1, pp. 291–297, 2014.

[233] J.-Y. Chien, W.-Y. Chiu, S.-T. Chien, C.-J. Chiang, C.-J. Yu, and P.-R. Hsueh, "Mutations in gyra and gyrb among fluoroquinolone-and multidrug-resistant mycobacterium tuberculosis isolates," *Antimicrobial agents and chemotherapy*, vol. 60, no. 4, pp. 2090–2096, 2016.

[234] S. Y. Wong, J. S. Lee, H. K. Kwak, L. E. Via, H. I. Boshoff, and C. E. Barry III, "Mutations in gidb confer low-level streptomycin resistance in mycobacterium tuberculosis," *Antimicrobial agents and chemotherapy*, vol. 55, no. 6, pp. 2515–2522, 2011.

[235] F. Brossier, N. Veziris, C. Truffot-Pernot, V. Jarlier, and W. Sougakoff, "Molecular investigation of resistance to the antituberculous drug ethionamide in multidrug-resistant clinical isolates of mycobacterium tuberculosis," *Antimicrobial agents and chemotherapy*, vol. 55, no. 1, pp. 355–360, 2011.

[236] Y. Shinde, I. Ahmad, S. Surana, and H. Patel, "The mur enzymes chink in the armour of mycobacterium tuberculosis cell wall," *European Journal of Medicinal Chemistry*, vol. 222, p. 113568, 2021.

[237] M. Al-Saeedi and S. Al-Hajoj, "Diversity and evolution of drug resistance mechanisms in mycobacterium tuberculosis," *Infection and drug resistance*, pp. 333–342, 2017.

[238] J. Kardan-Yamchi, H. Kazemian, M. Haeili, A. A. Harati, S. Amini, and M. M. Feizabadi, "Expression analysis of 10 efflux pump genes in multidrug-resistant and extensively drug-resistant mycobacterium tuberculosis clinical isolates," *Journal of Global Antimicrobial Resistance*, vol. 17, pp. 201–208, 2019.

[239] M. T. Zaw, N. A. Emran, and Z. Lin, "Mutations inside rifampicin-resistance determining region of rpob gene associated with rifampicin-resistance in mycobacterium tuberculosis," *Journal of infection and public health*, vol. 11, no. 5, pp. 605–610, 2018.

[240] L. D. Forsman, C. Giske, J. Bruchfeld, T. Schon, P. Jureen, and K. Angeby, "Meropenem-clavulanate has high in vitro activity against multidrug-resistant mycobacterium tuberculosis," *The International Journal of Mycobacteriology*, vol. 4, no. Suppl 1, pp. S80–S81, 2015.

[241] H. Hiasa, "The glu-84 of the parc subunit plays critical roles in both topoisomerase iv- quinolone and topoisomerase iv- dna interactions," *Biochemistry*, vol. 41, no. 39, pp. 11779–11785, 2002.

[242] D. Sheng, X. Chen, Y. Li, J. Wang, L. Zhuo, and Y. Li, "Parc, a new partitioning protein, is necessary for the active form of para from myxococcus pmf1 plasmid," *Frontiers in Microbiology*, vol. 11, p. 623699, 2021.

[243] J. R. McCann, J. A. McDonough, M. S. Pavelka, and M. Braunstein, "B-lactamase can function as a reporter of bacterial protein export during mycobacterium tuberculosis infection of host cells," *Microbiology*, vol. 153, no. 10, pp. 3350–3359, 2007.

[244] T. Schlothauer, A. Mogk, D. A. Dougan, B. Bukau, and K. Turgay, "Meca, an adaptor protein necessary for clpc chaperone activity," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2306–2311, 2003.

[245] R. Garcia-Castellanos, G. Mallorqui-Fernandez, A. Marrero, J. Potempa, M. Coll, and F. X. Gomis-Ruth, "On the transcriptional regulation of methicillin resistance: Meci repressor in complex with its operator," *Journal of Biological Chemistry*, vol. 279, no. 17, pp. 17888–17896, 2004.

[246] M. C. J. Blokpoel, H. N. Murphy, R. O'Toole, S. Wiles, E. S. C. Runn, G. R. Stewart, D. B. Young, and B. D. Robertson, "Tetracycline-inducible gene regulation in mycobacteria," *Nucleic Acids Research*, vol. 33, pp. e22–e22, 01 2005.

[247] R. S. Singh, K. Chauhan, and J. F. Kennedy, "A panorama of bacterial inulinases: production, purification, characterization and industrial applications," *International journal of biological macromolecules*, vol. 96, pp. 312–322, 2017.

[248] A. L. Stern, S. E. Van der Verren, J. Nasvall, H. Gutierrez-de Teran, M. Selmer, *et al.*, "Structural mechanism of aada, a dual-specificity aminoglycoside adenylyltransferase from salmonella enterica," *Journal of Biological Chemistry*, vol. 293, no. 29, pp. 11481–11490, 2018.

[249] S. Gu, S. Rumpel, J. Zhou, J. Strotmeier, H. Bigalke, K. Perry, C. B. Shoemaker, A. Rummel, and R. Jin, "Botulinum neurotoxin is shielded by ntnha in an interlocked complex," *Science*, vol. 335, no. 6071, pp. 977–981, 2012.

[250] J. Zhu, M. He, W. Xu, Y. Li, R. Huang, S. Wu, and H. Niu, "Development of tem-1 b-lactamase based protein translocation assay for identification of anaplasma phagocytophilum type iv secretion system effector proteins," *Scientific Reports*, vol. 9, no. 1, p. 4235, 2019.

[251] J. C. Morse, D. Girodat, B. J. Burnett, M. Holm, R. B. Altman, K. Y. Sanbonmatsu, H.-J. Wieden, and S. C. Blanchard, "Elongation factor-tu can repetitively engage aminoacyl-trna within the ribosome during the proofreading stage of trna selection," *Proceedings of the National Academy of Sciences*, vol. 117, no. 7, pp. 3610–3620, 2020.

[252] E. Dzyubak and M.-N. F. Yap, "The expression of antibiotic resistance methyltransferase correlates with mrna stability independently of ribosome stalling," *Antimicrobial agents and chemotherapy*, vol. 60, no. 12, pp. 7178–7188, 2016.

[253] M. J. Nasiri, H. Dabiri, D. Darban-Sarokhalil, M. Rezadehbashi, and S. Zamani, "Prevalence of drug-resistant tuberculosis in iran: systematic review and meta-analysis," *American journal of infection control*, vol. 42, no. 11, pp. 1212–1218, 2014.

[254] C. C. Onyedum, I. Alobu, and K. N. Ukwaja, "Prevalence of drug-resistant tuberculosis in nigeria: A systematic review and meta-analysis," *PloS one*, vol. 12, no. 7, p. e0180996, 2017.

[255] G. Bahizi, R. K. Majwala, S. Kisaka, A. Nyombi, K. Musisi, B. Kwesiga, L. Bulage, A. R. Ario, and S. Turyahabwe, "Epidemiological profile of patients with rifampicin-resistant tuberculosis: an analysis of the uganda national tuberculosis reference laboratory surveillance data, 2014–2018," *Antimicrobial Resistance, Infection Control*, vol. 10, no. 1, p. 76, 2021.

[256] R. P. Jaktaji and E. Mohiti, "Study of mutations in the dna gyrase gyra gene of escherichia coli," *Iranian journal of pharmaceutical research: IJPR*, vol. 9, no. 1, p. 43, 2010.

[257] A. Ghosh, S. N, and S. Saha, "Survey of drug resistance associated gene mutations in mycobacterium tuberculosis, eskape and other bacterial species," *Scientific reports*, vol. 10, no. 1, p. 8957, 2020.

[258] K. Hosokawa, N.-H. Park, T. Inaoka, Y. Itoh, and K. Ochi, "Streptomycin-resistant (rpsl) or rifampicin-resistant (rpob) mutation in pseudomonas putida kh146-2 confers enhanced tolerance to organic chemicals," *Environmental microbiology*, vol. 4, no. 11, pp. 703–712, 2002.

[259] P. Praest, R. Luteijn, I. Brak-Boer, J. Lanfermeijer, H. Hoelen, L. Ijgosse, A. Costa, R. Gorham Jr, R. Lebbink, and E. Wiertz, "The influence of tap1 and tap2 gene polymorphisms on tap function and its inhibition by viral immune evasion proteins," *Molecular immunology*, vol. 101, pp. 55–64, 2018.

[260] P. Purkan, I. Ihsanawati, D. Natalia, Y. Syah, D. Retnoningrum, and I. Siswanto, "Molecular analysis of katg encoding catalase-peroxidase from clinical isolate of isoniazid-resistant mycobacterium tuberculosis," *Journal of medicine and life*, vol. 11, no. 2, p. 160, 2018.

[261] R. Wang, K. Li, J. Yu, J. Deng, and Y. Chen, "Mutations of folc cause increased susceptibility to sulfamethoxazole in mycobacterium tuberculosis," *Scientific Reports*, vol. 11, no. 1, p. 1352, 2021.

[262] Y. Sasaki, Y. Ito, and T. Sasaki, "Thya as a selection marker in construction of food-grade host-vector and integration systems for streptococcus thermophilus," *Applied and environmental microbiology*, vol. 70, no. 3, pp. 1858–1864, 2004.

[263] S. Sinha-Ray and A. Ali, "Mutation in flra and msha genes of vibrio cholerae inversely involved in vps-independent biofilm driving bacterium toward nutrients in lake water," *Frontiers in Microbiology*, vol. 8, p. 1770, 2017.

[264] S. R. Luckner, C. A. Machutta, P. J. Tonge, and C. Kisker, "Crystal structures of mycobacterium tuberculosis kasa show mode of action within cell wall biosynthesis and its inhibition by thiolactomycin," *Structure*, vol. 17, no. 7, pp. 1004–1013, 2009.

[265] S. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. Gordon, K. Eiglmeier, S. Gas, C. Barry Iii, *et al.*, "Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence," *Nature*, vol. 396, no. 6707, pp. 190–190, 1998.

[266] A. Van den Broeke, M. Van Poucke, A. Marcos-Carcavilla, K. Hugot, H. Hayes, M. Bertaud, A. Van Zeveren, and L. J. Peelman, "Characterization of the ovine ribosomal protein sa gene and its pseudogenes," *BMC genomics*, vol. 11, no. 1, pp. 1–12, 2010.

[267] J. Perdigao, R. Macedo, D. Machado, C. Silva, L. Jordao, I. Couto, M. Viveiros, and I. Portugal, "Gidb mutation as a phylogenetic marker for q1 cluster mycobacterium tuberculosis isolates and intermediate-level streptomycin resistance determinant in lisbon, portugal," *Clinical microbiology and infection*, vol. 20, no. 5, pp. O278–O284, 2014.

[268] P. K. Anand, A. Kumar, A. Saini, and J. Kaur, "Mutation in eth a protein of mycobacterium tuberculosis conferred drug tolerance against enthinoamide in mycobacterium smegmatis mc2155," *Computational Biology and Chemistry*, vol. 98, p. 107677, 2022.

[269] J. Liu, W. Shi, S. Zhang, X. Hao, D. A. Maslov, K. V. Shur, O. B. Bekker, V. N. Danilenko, and Y. Zhang, "Mutations in efflux pump rv1258c (tap) cause resistance to pyrazinamide, isoniazid, and streptomycin in m. tuberculosis," *Frontiers in Microbiology*, vol. 10, p. 216, 2019.

[270] G. Brandis, M. Wrande, L. Liljas, and D. Hughes, "Fitness-compensatory mutations in rifampicin-resistant rna polymerase," *Molecular microbiology*, vol. 85, no. 1, pp. 142–151, 2012.

[271] M. De Vos, B. Muller, S. Borrell, P. Black, P. Van Helden, R. Warren, S. Gagneux, and T. Victor, "Putative compensatory mutations in the rpoc gene of rifampin-resistant mycobacterium tuberculosis are associated with ongoing transmission," *Antimicrobial agents and chemotherapy*, vol. 57, no. 2, pp. 827–832, 2013.

[272] A. S. Khan, J. E. Phelan, M. T. Khan, S. Ali, M. Qasim, G. Napier, S. Campino, S. Ahmad, O. Cabral-Marques, S. Zhang, *et al.*, "Characterization of rifampicin-resistant mycobacterium tuberculosis in khyber pakhtunkhwa, pakistan," *Scientific reports*, vol. 11, no. 1, p. 14194, 2021.

[273] I. Comas, S. Borrell, A. Roetzer, G. Rose, B. Malla, M. Kato-Maeda, J. Galagan, S. Niemann, and S. Gagneux, "Whole-genome sequencing of rifampicin-resistant mycobacterium tuberculosis strains identifies compensatory mutations in rna polymerase genes," *Nature genetics*, vol. 44, no. 1, pp. 106–110, 2012.

[274] Y. J. Yun, J. S. Lee, J. C. Yoo, E. Cho, D. Park, Y.-H. Kook, and K. H. Lee, "Patterns of rpoc mutations in drug-resistant mycobacterium tuberculosis isolated from patients in south korea," *Tuberculosis and Respiratory Diseases*, vol. 81, no. 3, pp. 222–227, 2018.

Appendix A

.1 The list of the institutions sharing data on TB portals.

1. Scientific Research Institute of Lung Diseases, Ministry of Health, Baku Azerbaijan
2. The United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk
3. Republican Research and Practical Centre for Pulmonology and Tuberculosis, Minsk Belarus
4. Shenzhen Center for Chronic Disease Control, Shenzhen,China
5. Fondation Congolaise pour la Recherche Medicale Republic of the Congo
6. National Center for Tuberculosis and Lung Diseases, Ministry of Health, TBilisi Georgia
7. Manipal Academy of Higher Education, Karnataka India
8. National Science Center of Phthisiopulmonology, Almaty Kazakhstan
9. National Tuberculosis Center, Bishkek Kyrgyzstan
10. University Clinical Research Center, Bamako Mali
11. Universidad Autónoma de Nuevo León, Monterrey Mexico

12. Phthisiopneumology Institute, Chisinau Moldova

13. University of Ibadan, Ibadan Nigeria,

14. Marius Nasta Pneumophtisiology Institute, Bucuresti Romania

15. Hospital Center University De Fann, Senegal Senegal

16. Perinatal HIV Research Unit, Johannesburg South Africa

17. Kharkiv National Medical University, Kharkiv Ukraine

18. National Lung Hospital, Hanoi Vietnam

Appendix B

.2 List of Demographic Features and Genomic Information

Sr. No.	Feature/ Attribute	Description	Feature Selection
1	condition id	Identification No. of patient throughout Dataset	
2	gender	Demographic Feature	YES
3	age of onset	Demographic Feature	YES
4	type of resistance	Demographic Feature	YES
5	outcome	Treatment Results	YES
6	specimen id	ID used by the specific source	
7	specimen identifier	ID used by the specific source	
8	specimen collection date	ID used by the specific source	
9	specimen collection site	Type of Sample	
10	sra id	Record for entry in the Read Archive (SRA) database.	
11	ncbi sra	Record identifier assigned by NCBI for entry in the SRA database	
12	ncbi sourceorganism	Record identifier for Mycobacterium tuberculosis organism sample in the NCBI SRA database	

Sr. No.	**Feature/ Attribute**	**Description**	**Feature Selection**
13	ncbi bioproject	Record identifier assigned by NCBI for entry in the BioProject database.	
14	ncbi biosample	Record identifier assigned by NCBI for entry in the BioSample database.	
15	lineage	Localization/ Classification	YES
16	sit designation	Demographic Feature	
17	gene snp mutations	snp Identified	
18	high confidence snp mutations	Binary data	
19	hain snp mutations	SNP mutations identified using Line-Probe Assay (LPA) test systems for Hain Lifescience molecular genetic diagnosis of mycobacteria.	
20	Genexpert snp mutations	SNP mutations identified using Xpert MTB / RIF test	

Appendix C

.3 Codes

1. Explanatory Data Analysis

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_theme(style="darkgrid")
print("Done")

data = pd.read_csv('TB_Portals_Genomics_October_2021.csv')
data.info()
df=data
sns.countplot(x='type_of_resistance', hue='outcome', data=df, palette="crest")
plt.legend(bbox_to_anchor=(1.02, 1), loc='upper left', borderaxespad=0)
plt.xticks(rotation=90)

plt.title('Type of Resistance vs Outcome Status')
plt.xlabel('Type of resistance')
plt.ylabel('Count')
# Save the figure
plt.savefig("barplot_outcome_resistance(2).png", dpi=300, bbox_inches='tight')

# Show the plot
plt.show()

df = df[df["type_of_resistance"] != "Sensitive"]

df.info()

df = df[df["outcome"] != "Cured"]

df.info()

df.to_csv("TB_subset1.csv")

data=df

data.head()

data.tail()

data.info

plt.figure(figsize=(10,10))

cmap=sns.color_palette("crest", as_cmap=True)
```

```
sns.countplot(x='gender', data=df, palette="crest")
plt.xlabel('Gender')
plt.ylabel('Count')
plt.title('Gender Distribution')


# Save the figure
#plt.savefig("barplot_gender_distribution.png", dpi=500)

# Show the plot
plt.show()

sns.histplot(data = data
          ,x = 'age_of_onset'
          ,color = 'SteelBlue'
          ,bins=5)

plt.xlabel('Age')
plt.ylabel('Count')
plt.title('Age of Onset')

# save the fig
plt.savefig('histogram_AGE.png', dpi=500)

# show the plot
plt.show()

sns.countplot(x='type_of_resistance', data=df, palette="crest")

plt.xlabel('Resistance')
plt.ylabel('Count')
plt.title('Type of Resistance')

# Save the figure
plt.savefig("barplot_resistance.png", dpi=300)

# Show the plot
plt.show()

plt.figure(figsize=(10,10))

sns.countplot(x='outcome', data=df, palette="crest")
plt.xticks(rotation=90)

plt.xlabel('Outcomes')
plt.ylabel('Count')
plt.title('Outcome')
# Save the figure
fig = plt.gcf()
fig.set_size_inches(10,8)
plt.savefig("barplot_outcome.png", dpi=300, bbox_inches='tight')
```

```
# Show the plot
plt.show()

#plt.figure(figsize=(10,10))

sns.countplot(x='ncbi_sourceorganism', data=df, palette="crest")
#plt.xticks(rotation=90)

#plt.xlabel('Source Organism')
#plt.ylabel('Count')
#plt.title('NCBI Source Organism Distribution')

# Save the figure
#fig = plt.gcf()
#fig.set_size_inches(10,8)
#plt.savefig("barplot_ncbi_source_organism.png", dpi=500, bbox_inches='tight')

plt.figure(figsize=(10,10))

sns.countplot(x='lineage', data=df, palette="crest")
plt.xticks(rotation=90)

plt.xlabel('Lineage')
plt.ylabel('Count')
plt.title('Lineage Distribution')

# Save the figure
fig = plt.gcf()
fig.set_size_inches(10,8)
plt.savefig("barplot_lineage.png", dpi=500, bbox_inches='tight')

# Show the plot
plt.show()

sns.countplot(x='gender', hue='outcome', data=df, palette="crest")

plt.title('Gender vs Outcome Status')
plt.xlabel('Gender')
plt.ylabel('Count')


# Save the figure
plt.savefig("barplot_gender_outcome.png", dpi=300)

# Show the plot
plt.show()

sns.countplot(x='gender', hue='type_of_resistance', data=df, palette="crest")

plt.title('Gender vs Type of Resistance')
plt.xlabel('Gender')
plt.ylabel('Count')
```

```
# Save the figure
plt.savefig("barplot_gender_resistance.png", dpi=300)

# Show the plot
plt.show()

sns.countplot(x='lineage', hue='gender', data=df, palette="crest")
plt.legend(loc='upper right')

plt.xticks(rotation=90)
plt.title('Gender vs Lineage Status')
plt.xlabel('Lineage')
plt.ylabel('Count')


# Save the figure
fig = plt.gcf()
fig.set_size_inches(10,8)
plt.savefig("barplot_gender_lineage.png", dpi=300, bbox_inches='tight')

# Show the plot
plt.show()

sns.boxplot(data=df, x="age_of_onset", y="outcome", palette="crest")
plt.title('Age of Onset vs Outcome')
plt.xlabel('Age')
plt.ylabel('Outcome')

fig = plt.gcf()
fig.set_size_inches(10,8)
plt.savefig("boxplot_age_outcome.png", dpi=300, bbox_inches='tight')

sns.boxplot(data=df, x="age_of_onset", y="type_of_resistance", palette="crest")
plt.title('Age of Onset vs Type of Resistance')
plt.xlabel('Age')
plt.ylabel('Resistance')

plt.savefig("boxplot_age_resistance.png")

sns.boxplot(data=df, x="age_of_onset", y="lineage", palette="crest")

plt.title('Age of Onset vs Lineage')
plt.xlabel('Age')
plt.ylabel('Lineage')

fig = plt.gcf()
fig.set_size_inches(10,8)
plt.savefig("boxplot_age_lineage.png", dpi=300, bbox_inches='tight')

sns.countplot(x='type_of_resistance', hue='outcome', data=df, palette="crest")

plt.title('Type of Resistance vs Outcome Status')
```

```
plt.xlabel('Type of resistance')
plt.ylabel('Count')


# Save the figure
plt.savefig("barplot_outcome_resistance.png", dpi=300)

# Show the plot
plt.show()

sns.countplot(x='lineage', hue='outcome', data=df, palette="crest")

plt.legend(bbox_to_anchor=(1.02, 1), loc='upper right', borderaxespad=0)

plt.title('Lineage vs Outcome Status')
plt.xlabel('Lineage')
plt.ylabel('Count')
plt.xticks(rotation=90)


fig = plt.gcf()
fig.set_size_inches(10,8)
plt.savefig("barplot_outcome_lineage.png", dpi=300, bbox_inches='tight')


# Show the plot
plt.show()

sns.countplot(x='lineage', hue='type_of_resistance', data=df, palette="crest")

plt.legend(bbox_to_anchor=(1.02, 1), loc='upper right', borderaxespad=0)

plt.xticks(rotation=90)
plt.title('Type of Resistance w.r.t Lineage')
plt.xlabel('Lineage')
plt.ylabel('Count')

fig = plt.gcf()
fig.set_size_inches(10,8)
plt.savefig("barplot_resistance_lineage.png", dpi=500, bbox_inches='tight')
```

2. Association rule mining

```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Read the CSV file into a list of transactions
transactions = []
with open('/home/DST sheet-R replaced with drugname - S and I removed.csv', 'r') as f:
    for line in f:
        items = line.strip().split(',')  # Adjust the delimiter if needed
```

```
        transactions.append(items)

# Convert the list of transactions to a DataFrame
te = TransactionEncoder()
te_ary = te.fit_transform(transactions)
df = pd.DataFrame(te_ary, columns=te.columns_)

# Mine frequent itemsets using Apriori algorithm
min_support = 0.1  # Adjust the minimum support threshold as per your requirement
frequent_itemsets = apriori(df, min_support=min_support, use_colnames=True)

# Convert frozensets to lists
frequent_itemsets['itemsets'] = frequent_itemsets['itemsets'].apply(lambda x: list(x))

# Generate association rules
min_confidence = 0.8  # Adjust the minimum confidence threshold as per your requirement
rules = association_rules(frequent_itemsets, metric='confidence', min_threshold=min_confidence)

# Convert frozensets to lists
rules['antecedents'] = rules['antecedents'].apply(lambda x: ', '.join(list(x)))
rules['consequents'] = rules['consequents'].apply(lambda x: ', '.join(list(x)))

# Save association rules to a CSV file
rules.to_csv('/home/association_ruleswithformating.csv', index=False)
```

3. Ariba input

1. Common Clusters

```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": 21,
   "id": "055756a9",
   "metadata": {},
   "outputs": [],
   "source": [
    "import pandas as pd\n",
    "\n",
    "# List of input file paths\"\n",
    "input_files = [\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_1.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_2.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_3.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_4.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_5.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_6.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_7.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_8.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_9.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_10.csv\",
```

```
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_11.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_12.csv\"]\n",
   "\n",
   "# List to store cluster column data from all files\n",
   "cluster_data = []\n",
   "\n",
   "# Read cluster column data from each file\n",
   "for file in input_files:\n",
   "    df = pd.read_csv(file, delimiter='\\t')\n",
   "    cluster_data.append(df['cluster'].tolist())\n",
   "#print(cluster_data)\n",
   "#Combine cluster column data into a single list\n",
   "combined_clusters = [cluster for clusters in cluster_data for cluster in clusters]\n",
   "#print(combined_clusters)\n",
   "\n",
   "\n",
   "# Identify unique clusters\n",
   "unique_clusters = set(combined_clusters)\n",
   "#print(unique_clusters)\n",
   "\n",
   "# # Create a DataFrame to store the common cluster rows\n",
   "common_cluster_df = pd.DataFrame()\n",
   "\n",
   "# # Compare files based on clusters and populate common_cluster_df\n",
   "for cluster in unique_clusters:\n",
   "    cluster_rows = []\n",
   "    for i, file in enumerate(input_files):\n",
   "        df = pd.read_csv(file, delimiter='\\t')\n",
   "        cluster_rows.append(df[df['cluster'] == cluster])\n",
   "    common_cluster_rows = pd.concat(cluster_rows, ignore_index=True)\n",
   "    common_cluster_df = pd.concat([common_cluster_df, common_cluster_rows])\n",
   "   # common_cluster_df = common_cluster_df.append(common_cluster_rows)\n",
   "\n",
   "\n",
   "# Write common cluster DataFrame to a file\n",
   "common_cluster_df.to_csv(\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/common_clusters.csv\",
index=False)\n",
   "# combined_clusters = pd.concat(cluster_data, ignore_index=True).tolist()\n",
   "# print(combined_clusters)"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 11,
  "id": "1620cfc5",
  "metadata": {},
  "outputs": [
   {
    "name": "stdout",
    "output_type": "stream",
    "text": [
     "[['CDD_-', 'CDD_-', 'CDD_-', 'cdeA', 'cdeA', 'tet_32_', 'tet_32_', 'tet_32_', 'tet_32_', 'tet_32_', 'tet_32_',
'tet_O_32_O_', 'tet_O_32_O_', 'tet_O_32_O_']]\n"
```

```
   ]
   }
  ],
  "source": [
   "import pandas as pd\n",
   "\n",
   "# List of input file paths\"\n",
   "input_files = [\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_1.csv\"]\n",
   "          #\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_2.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_3.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_4.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_5.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_6.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_7.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_8\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_9.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_10.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_11.csv\",
\"/home/Anam/Desktop/ariba/PRJ_DB/novel_variants/novel_variant_12.csv\"]\n",
   "\n",
   "# List to store cluster column data from all files\n",
   "cluster_data = []\n",
   "\n",
   "# Read cluster column data from each file\n",
   "for file in input_files:\n",
   "    df = pd.read_csv(file, delimiter='\\t')\n",
   "    cluster_data.append(df['cluster'].tolist())\n",
   "print(cluster_data)"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": null,
  "id": "2af78862",
  "metadata": {},
  "outputs": [],
  "source": []
 }
],
"metadata": {
 "kernelspec": {
  "display_name": "Python 3 (ipykernel)",
  "language": "python",
  "name": "python3"
 },
 "language_info": {
  "codemirror_mode": {
   "name": "ipython",
   "version": 3
  },
  "file_extension": ".py",
  "mimetype": "text/x-python",
  "name": "python",
```

```
  "nbconvert_exporter": "python",
  "pygments_lexer": "ipython3",
  "version": "3.10.9"
 }
},
"nbformat": 4,
"nbformat_minor": 5
}
```

2. Novel Variants

```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": 93,
   "id": "27247192",
   "metadata": {},
   "outputs": [],
   "source": [
    "import os\n",
    "import pandas as pd\n",
    "\n",
    "\n",
    "folder_path = \"/home/Anam/Desktop/ariba/PRJ_DB\"\n",
    "\n",
    "file_name = \"report.tsv\"\n",
    "file_list = []\n",
    "\n",
    "for root, dirs, files in os.walk(folder_path):\n",
    "    for file in files:\n",
    "        if file == file_name:\n",
    "            file_path = os.path.join(root, file)\n",
    "            file_list.append(file_path)\n",
    "#print(file_list)\n",
    "\n",
    "for file_path in file_list:\n",
    "    if file_path.endswith('.tsv'):\n",
    "        df = pd.read_csv(file_path, delimiter='\\t')\n",
    "        #print(df)\n",
    "        filtered_df = (df['gene'] == 1) & (df['known_var'] == '1')\n",
    "        known_variant = df[filtered_df]\n",
    "        #print(coding_gene)\n",
    "        output_file_path1 = os.path.join(os.path.dirname(file_path), \"known_variant.csv\")\n",
    "        print(output_file_path1)\n",
    "        known_variant.to_csv(output_file_path1, sep='\\t', index=False)\n",
    "        \n",
    "        \n",
    "        condition2 = (df['gene'] == 1) & (df['known_var'] == '0')\n",
    "        novel_variant = df[condition2]\n",
    "        #print(extracted_data2)\n",
```

```
  "       \n",
  "       output_file_path2 = os.path.join(os.path.dirname(file_path), \"novel_variant.csv\")\n",
  "       #print(output_file_path2)\n",
  "       novel_variant.to_csv(output_file_path2, sep='\\t', index=False)\n",
  "\n",
  "\n",
  "       \n",
  "\n",
  "           \n",
  "       "
 ]
},
{
 "cell_type": "code",
 "execution_count": 75,
 "id": "d93030b7",
 "metadata": {},
 "outputs": [],
 "source": [
  "\n"
 ]
},
{
 "cell_type": "code",
 "execution_count": null,
 "id": "58ad4f4c",
 "metadata": {},
 "outputs": [],
 "source": []
},
{
 "cell_type": "code",
 "execution_count": null,
 "id": "3690a8e0",
 "metadata": {},
 "outputs": [],
 "source": []
},
{
 "cell_type": "code",
 "execution_count": null,
 "id": "96e9d4bf",
 "metadata": {},
 "outputs": [],
 "source": []
}
],
"metadata": {
 "kernelspec": {
 "display_name": "Python 3 (ipykernel)",
 "language": "python",
 "name": "python3"
 },
```

```
  "language_info": {
   "codemirror_mode": {
    "name": "ipython",
    "version": 3
   },
   "file_extension": ".py",
   "mimetype": "text/x-python",
   "name": "python",
   "nbconvert_exporter": "python",
   "pygments_lexer": "ipython3",
   "version": "3.10.9"
  }
 },
 "nbformat": 4,
 "nbformat_minor": 5
}
```


Common Variants


```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": 18,
   "id": "2e99d90e",
   "metadata": {},
   "outputs": [
    {
     "name": "stdout",
     "output_type": "stream",
     "text": [
      "[]\n"
     ]
    },
    {
     "ename": "KeyError",
     "evalue": "'cluster'",
     "output_type": "error",
     "traceback": [
      "\u001b[0;31m---------------------------------------------------------------------------\u001b[0m",
      "\u001b[0;31mKeyError\u001b[0m                                Traceback (most recent call last)",
      "File \u001b[0;32m~/anaconda3/lib/python3.10/site-packages/pandas/core/indexes/base.py:3802\u001b[0m, in
\u001b[0;36mIndex.get_loc\u001b[0;34m(self, key, method, tolerance)\u001b[0m\n\u001b[1;32m
3801\u001b[0m \u001b[38;5;28;01mtry\u001b[39;00m:\n\u001b[0;32m-> 3802\u001b[0m
\u001b[38;5;28;01mreturn\u001b[39;00m
\u001b[38;5;28;43mself\u001b[39;49m\u001b[38;5;241;43m.\u001b[39;49m\u001b[43m_engine\u001b[49m\u00
1b[38;5;241;43m.\u001b[39;49m\u001b[43mget_loc\u001b[49m\u001b[43m(\u001b[49m\u001b[43mcasted_key\
u001b[49m\u001b[43m)\u001b[49m\n\u001b[1;32m   3803\u001b[0m \u001b[38;5;28;01mexcept\u001b[39;00m
\u001b[38;5;167;01mKeyError\u001b[39;00m \u001b[38;5;28;01mas\u001b[39;00m err:\n",
      "File \u001b[0;32m~/anaconda3/lib/python3.10/site-packages/pandas/_libs/index.pyx:138\u001b[0m, in
\u001b[0;36mpandas._libs.index.IndexEngine.get_loc\u001b[0;34m()\u001b[0m\n",
      "File \u001b[0;32m~/anaconda3/lib/python3.10/site-packages/pandas/_libs/index.pyx:165\u001b[0m, in
\u001b[0;36mpandas._libs.index.IndexEngine.get_loc\u001b[0;34m()\u001b[0m\n",
```

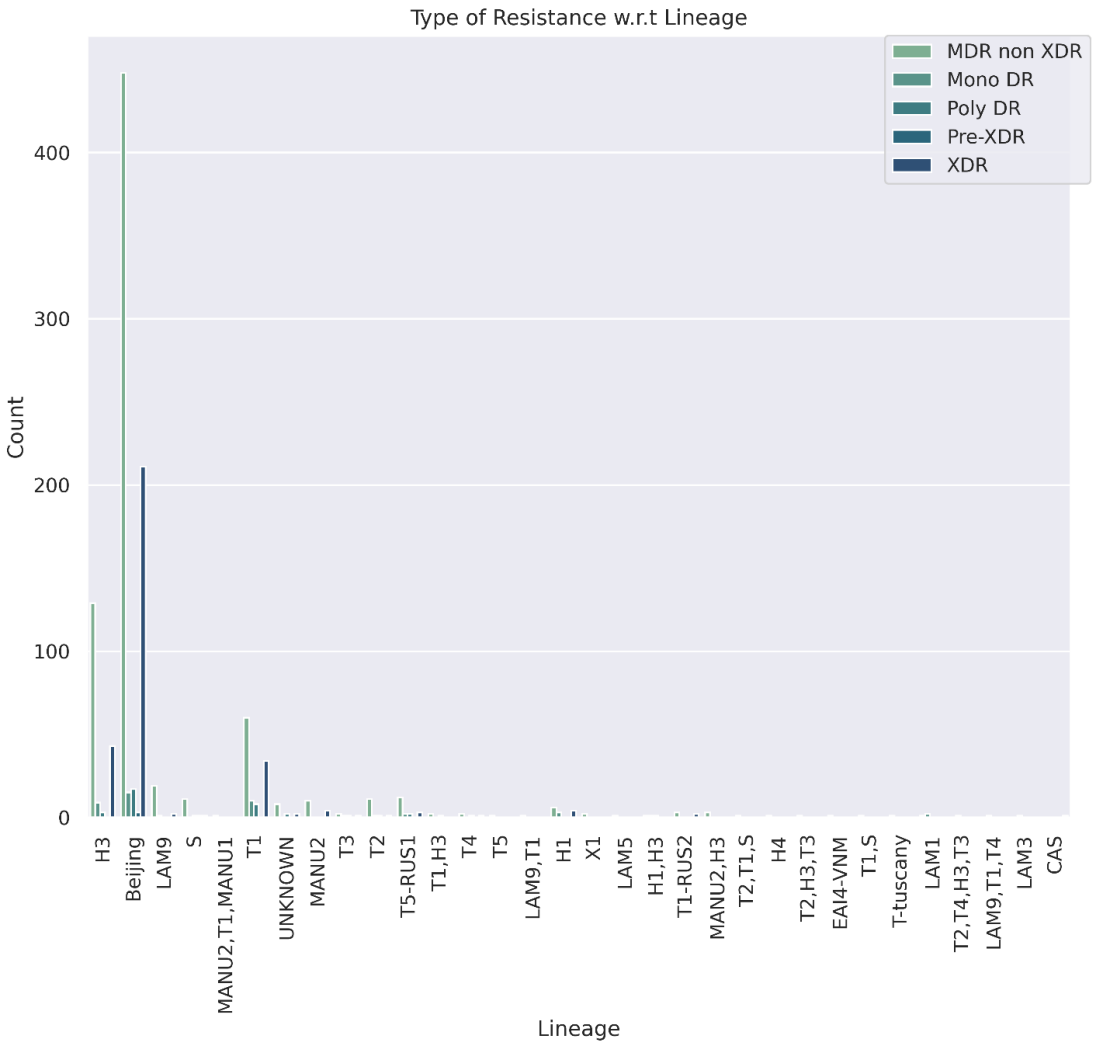
```
    "File \u001b[0;32mpandas/_libs/hashtable_class_helper.pxi:5745\u001b[0m, in
\u001b[0;36mpandas._libs.hashtable.PyObjectHashTable.get_item\u001b[0;34m()\u001b[0m\n",
    "File \u001b[0;32mpandas/_libs/hashtable_class_helper.pxi:5753\u001b[0m, in
\u001b[0;36mpandas._libs.hashtable.PyObjectHashTable.get_item\u001b[0;34m()\u001b[0m\n",
    "\u001b[0;31mKeyError\u001b[0m: 'cluster'",
    "\nThe above exception was the direct cause of the following exception:\n",
    "\u001b[0;31mKeyError\u001b[0m                             Traceback (most recent call last)",
    "Cell \u001b[0;32mIn[18], line 13\u001b[0m\n\u001b[1;32m    11\u001b[0m       df
\u001b[38;5;241m=\u001b[39m pd\u001b[38;5;241m.\u001b[39mread_csv(file_path,
delimiter\u001b[38;5;241m=\u001b[39m\u001b[38;5;124m\"\u001b[39m\u001b[38;5;130;01m\\t\u001b[39;00m\
u001b[38;5;124m\"\u001b[39m)\n\u001b[1;32m    12\u001b[0m
\u001b[38;5;66;03m#print(df)\u001b[39;00m\n\u001b[0;32m---> 13\u001b[0m
cluster_data\u001b[38;5;241m.\u001b[39mappend(\u001b[43mdf\u001b[49m\u001b[43m[\u001b[49m\u001b[38
;5;124;43m'\u001b[39;49m\u001b[38;5;124;43mcluster\u001b[39;49m\u001b[38;5;124;43m'\u001b[39;49m\u001
b[43m]\u001b[49m\u001b[38;5;241m.\u001b[39mtolist())\n\u001b[1;32m    15\u001b[0m combined_clusters
\u001b[38;5;241m=\u001b[39m [cluster \u001b[38;5;28;01mfor\u001b[39;00m clusters
\u001b[38;5;129;01min\u001b[39;00m cluster_data \u001b[38;5;28;01mfor\u001b[39;00m cluster
\u001b[38;5;129;01min\u001b[39;00m clusters]\n\u001b[1;32m    17\u001b[0m unique_clusters
\u001b[38;5;241m=\u001b[39m \u001b[38;5;28mset\u001b[39m(combined_clusters)\n",
    "File \u001b[0;32m~/anaconda3/lib/python3.10/site-packages/pandas/core/frame.py:3807\u001b[0m, in
\u001b[0;36mDataFrame.__getitem__\u001b[0;34m(self, key)\u001b[0m\n\u001b[1;32m   3805\u001b[0m
\u001b[38;5;28;01mif\u001b[39;00m
\u001b[38;5;28mself\u001b[39m\u001b[38;5;241m.\u001b[39mcolumns\u001b[38;5;241m.\u001b[39mnlevels
\u001b[38;5;241m>\u001b[39m \u001b[38;5;241m1\u001b[39m:\n\u001b[1;32m   3806\u001b[0m
\u001b[38;5;28;01mreturn\u001b[39;00m
\u001b[38;5;28mself\u001b[39m\u001b[38;5;241m.\u001b[39m_getitem_multilevel(key)\n\u001b[0;32m->
3807\u001b[0m indexer \u001b[38;5;241m=\u001b[39m
\u001b[38;5;28;43mself\u001b[39;49m\u001b[38;5;241;43m.\u001b[39;49m\u001b[43mcolumns\u001b[49m\u00
1b[38;5;241;43m.\u001b[39;49m\u001b[43mget_loc\u001b[49m\u001b[43m(\u001b[49m\u001b[43mkey\u001b[
49m\u001b[43m)\u001b[49m\n\n\u001b[1;32m   3808\u001b[0m \u001b[38;5;28;01mif\u001b[39;00m
is_integer(indexer):\n\u001b[1;32m   3809\u001b[0m     indexer \u001b[38;5;241m=\u001b[39m [indexer]\n",
    "File \u001b[0;32m~/anaconda3/lib/python3.10/site-packages/pandas/core/indexes/base.py:3804\u001b[0m, in
\u001b[0;36mIndex.get_loc\u001b[0;34m(self, key, method, tolerance)\u001b[0m\n\n\u001b[1;32m
3802\u001b[0m     \u001b[38;5;28;01mreturn\u001b[39;00m
\u001b[38;5;28mself\u001b[39m\u001b[38;5;241m.\u001b[39m_engine\u001b[38;5;241m.\u001b[39mget_loc(ca
sted_key)\n\n\u001b[1;32m   3803\u001b[0m \u001b[38;5;28;01mexcept\u001b[39;00m
\u001b[38;5;167;01mKeyError\u001b[39;00m \u001b[38;5;28;01mas\u001b[39;00m err:\n\u001b[0;32m->
3804\u001b[0m     \u001b[38;5;28;01mraise\u001b[39;00m \u001b[38;5;167;01mKeyError\u001b[39;00m(key)
\u001b[38;5;28;01mfrom\u001b[39;00m \u001b[38;5;21;01merr\u001b[39;00m\n\n\u001b[1;32m   3805\u001b[0m
\u001b[38;5;28;01mexcept\u001b[39;00m \u001b[38;5;167;01mTypeError\u001b[39;00m:\n\u001b[1;32m
3806\u001b[0m     \u001b[38;5;66;03m# If we have a listlike key, _check_indexing_error will
raise\u001b[39;00m\n\n\u001b[1;32m   3807\u001b[0m     \u001b[38;5;66;03m#  InvalidIndexError. Otherwise we
fall through and re-raise\u001b[39;00m\n\n\u001b[1;32m   3808\u001b[0m     \u001b[38;5;66;03m#  the
TypeError.\u001b[39;00m\n\n\u001b[1;32m   3809\u001b[0m
\u001b[38;5;28mself\u001b[39m\u001b[38;5;241m.\u001b[39m_check_indexing_error(key)\n",
    "\u001b[0;31mKeyError\u001b[0m: 'cluster'"
   ]
  }
 ],
 "source": [
  "import os\n",
  "import pandas as pd\n",
  "\n",
```

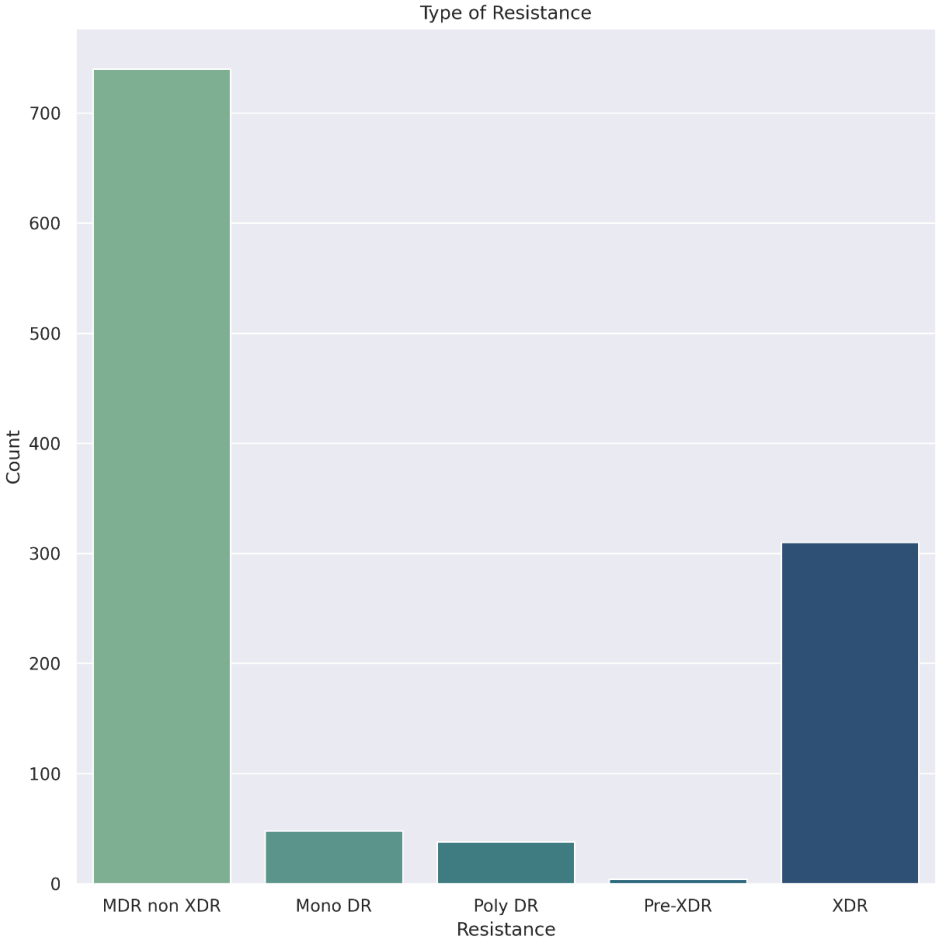
```
  "folder_path = \"/home/Anam/Desktop/xdr_novel_variant\"\n",
  "\n",
  "cluster_data = []\n",
  "print(cluster_data)\n",
  "for file_name in os.listdir(folder_path):\n",
  "    if file_name.endswith('.csv'):\n",
  "        file_path = os.path.join(folder_path, file_name)\n",
  "        df = pd.read_csv(file_path, delimiter=\"\\t\")\n",
  "        #print(df)\n",
  "        cluster_data.append(df['cluster'].tolist())\n",
  "\n",
  "combined_clusters = [cluster for clusters in cluster_data for cluster in clusters]\n",
  "\n",
  "unique_clusters = set(combined_clusters)\n",
  "\n",
  "common_cluster_df = pd.DataFrame()\n",
  "cluster_column_name = None  # Initialize the cluster column name\n",
  "for cluster in unique_clusters:\n",
  "    cluster_rows = []\n",
  "    for file_name in os.listdir(folder_path):\n",
  "        if file_name.endswith('.csv'):\n",
  "            file_path = os.path.join(folder_path, file_name)\n",
  "            df = pd.read_csv(file_path, delimiter=\"\\t\")\n",
  "            if cluster_column_name is None:\n",
  "                cluster_column_name = df.columns[6]  # Use the first column as cluster column\n",
  "            cluster_rows.append(df[df[cluster_column_name] == cluster])\n",
  "    common_cluster_rows = pd.concat(cluster_rows, ignore_index=True)\n",
  "    common_cluster_df = pd.concat([common_cluster_df, common_cluster_rows])\n",
  "\n",
  "common_cluster_df.to_csv(\"/home/Anam/Desktop/xdr_common1_clusters.csv\", index=False)\n"
 ]
},
{
 "cell_type": "code",
 "execution_count": null,
 "id": "5c0df94d",
 "metadata": {},
 "outputs": [],
 "source": []
},
{
 "cell_type": "code",
 "execution_count": null,
 "id": "fda90fcd",
 "metadata": {},
 "outputs": [],
 "source": []
},
{
 "cell_type": "code",
 "execution_count": null,
 "id": "608deda1",
 "metadata": {},
```

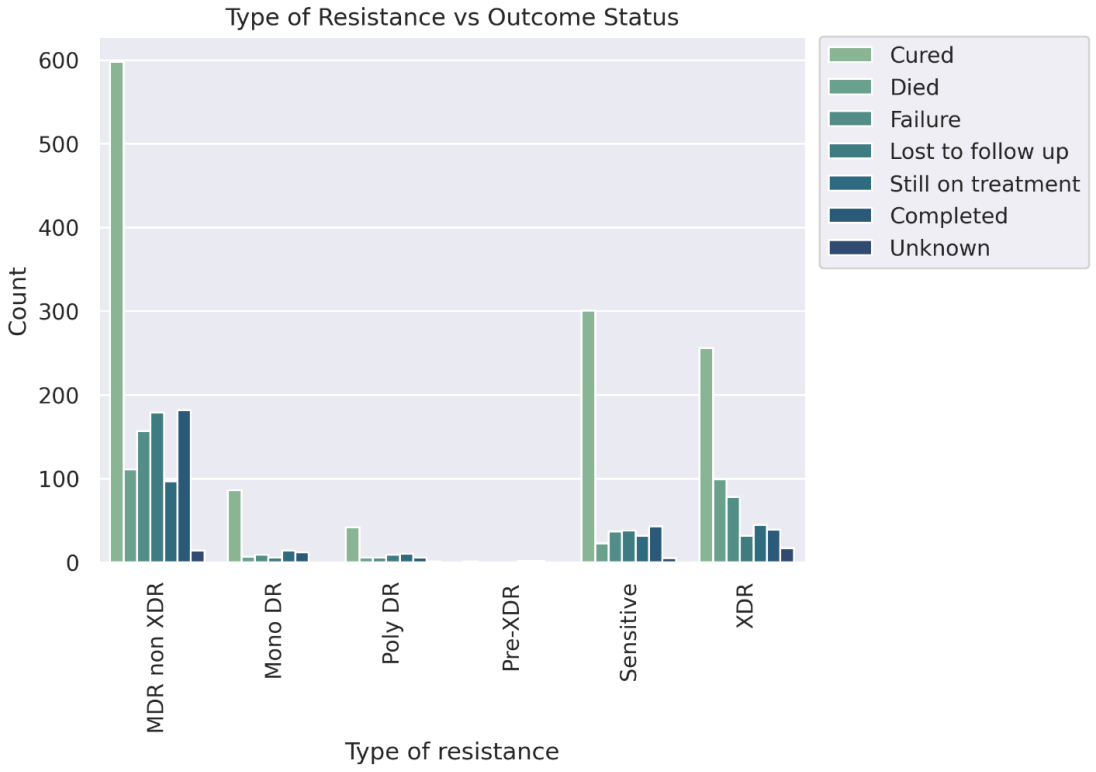
```
  "outputs": [],
  "source": []
 },
 {
  "cell_type": "code",
  "execution_count": null,
  "id": "1efe10dc",
  "metadata": {},
  "outputs": [],
  "source": []
 }
],
"metadata": {
 "kernelspec": {
  "display_name": "Python 3 (ipykernel)",
  "language": "python",
  "name": "python3"
 },
 "language_info": {
  "codemirror_mode": {
   "name": "ipython",
   "version": 3
  },
  "file_extension": ".py",
  "mimetype": "text/x-python",
  "name": "python",
  "nbconvert_exporter": "python",
  "pygments_lexer": "ipython3",
  "version": "3.10.9"
 }
},
"nbformat": 4,
"nbformat_minor": 5
}
```

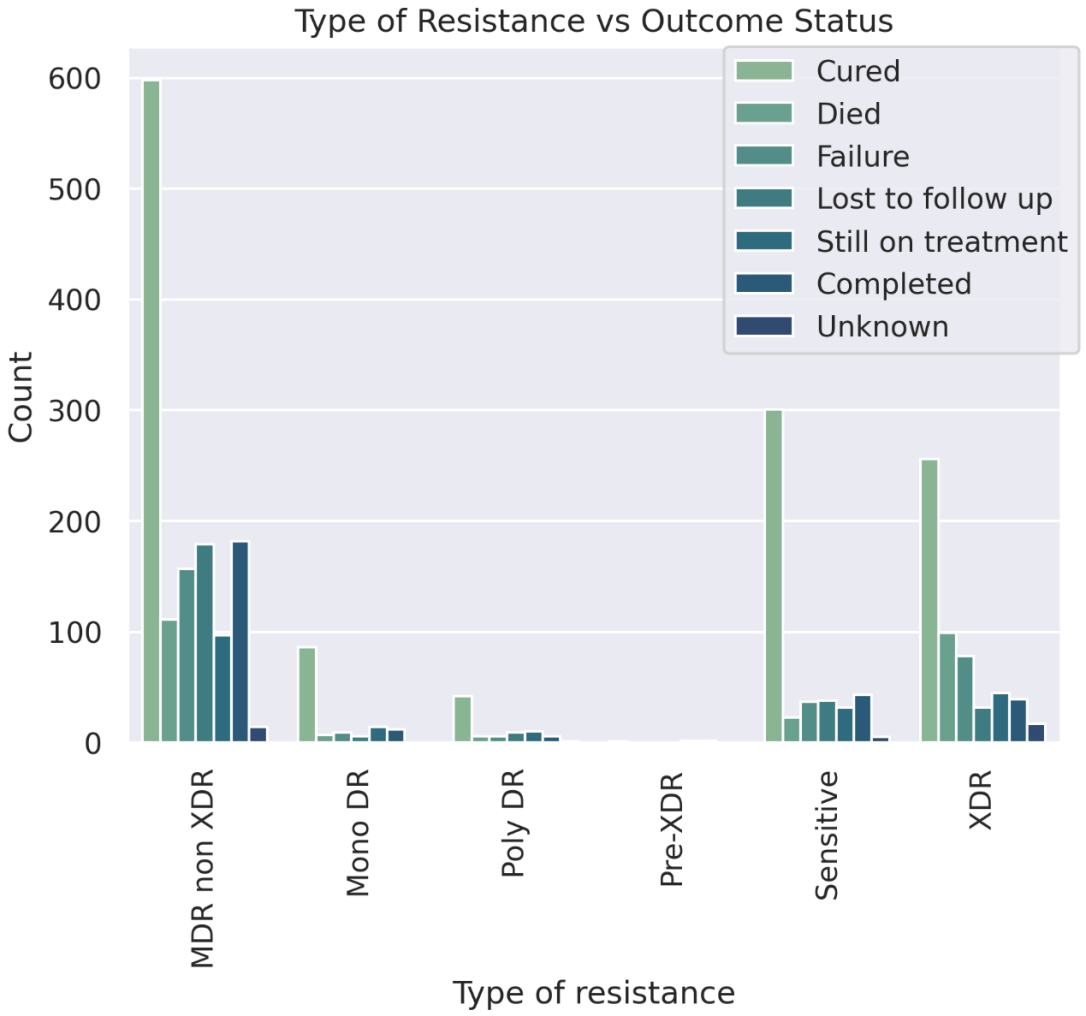
Appendix D

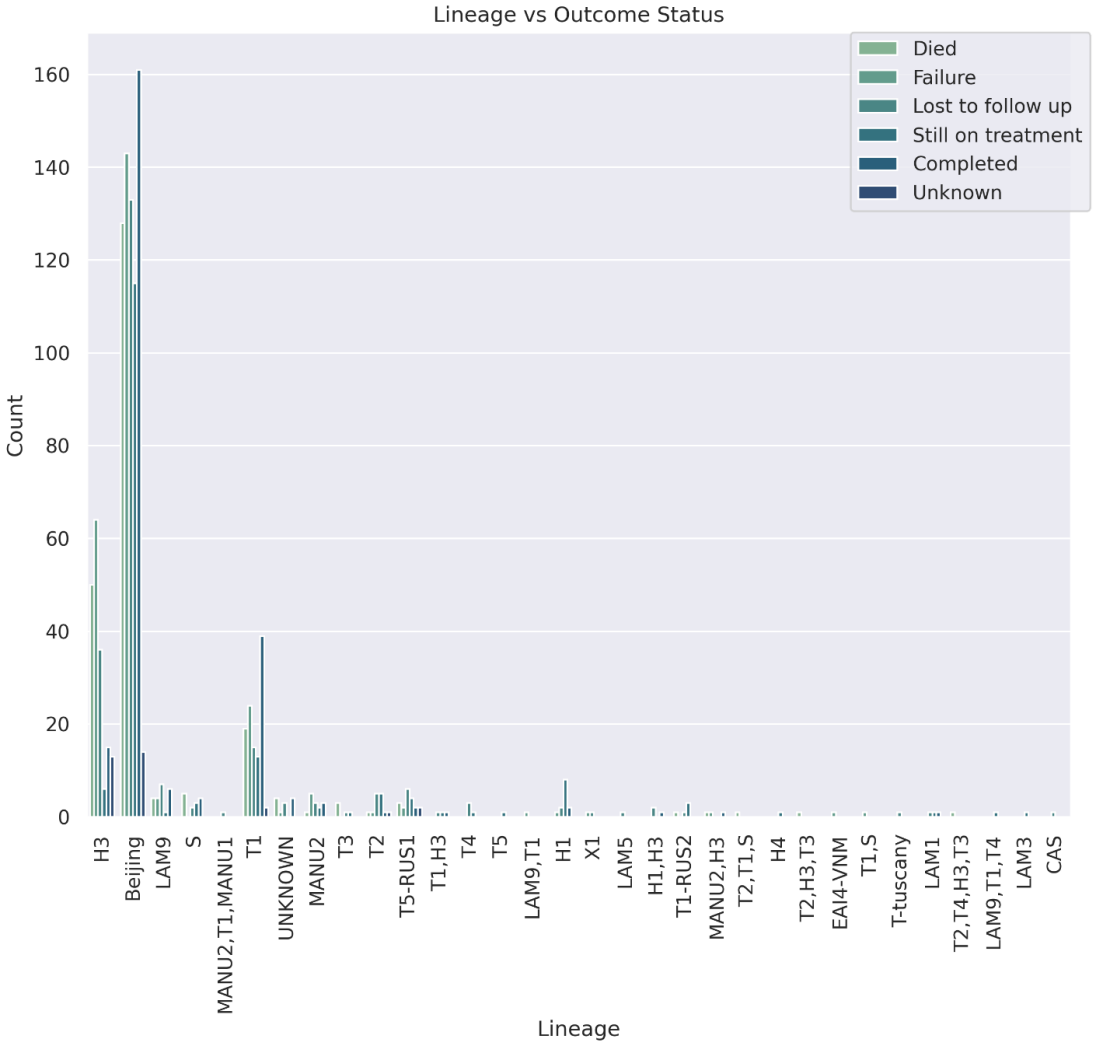
.4 TB Bar Graphs

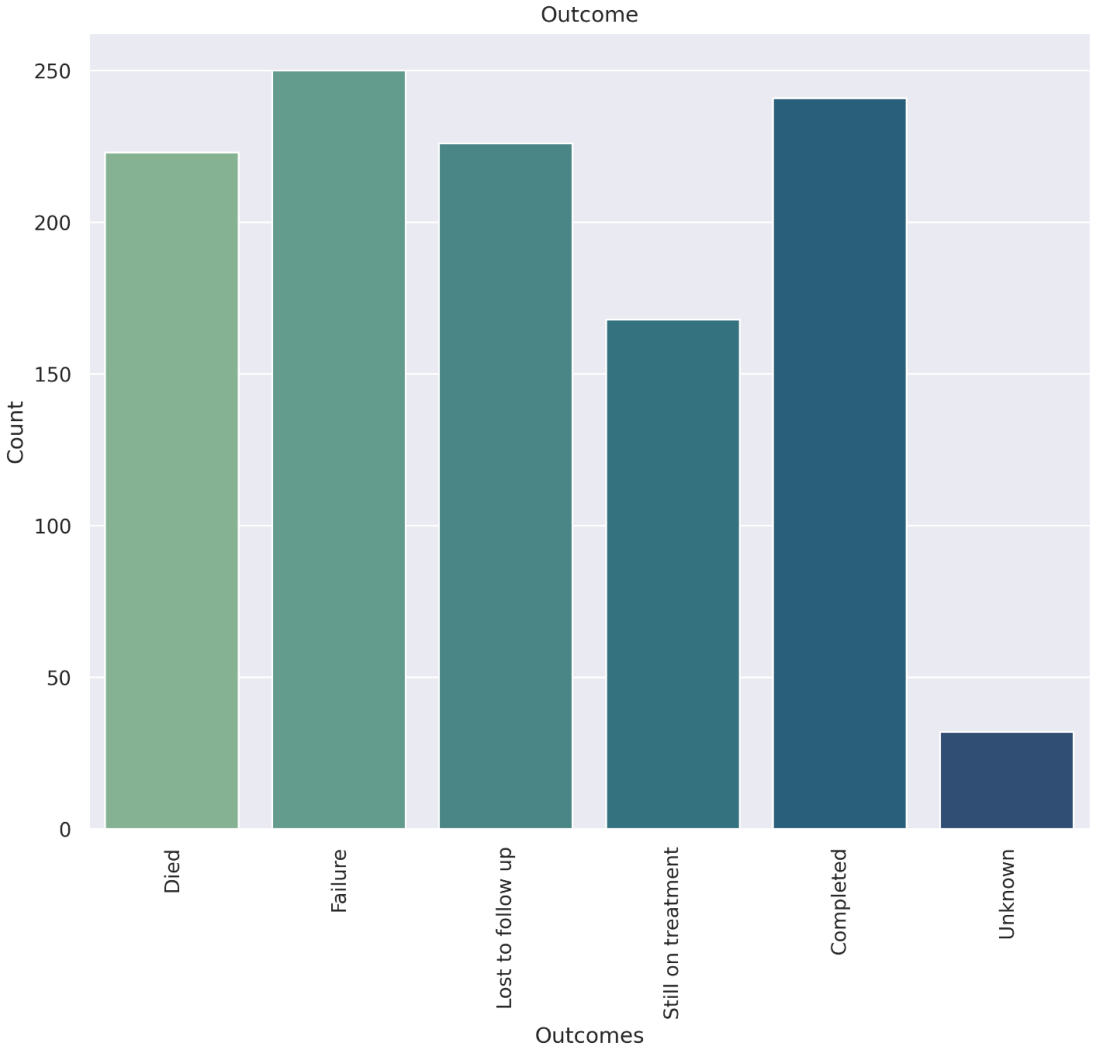


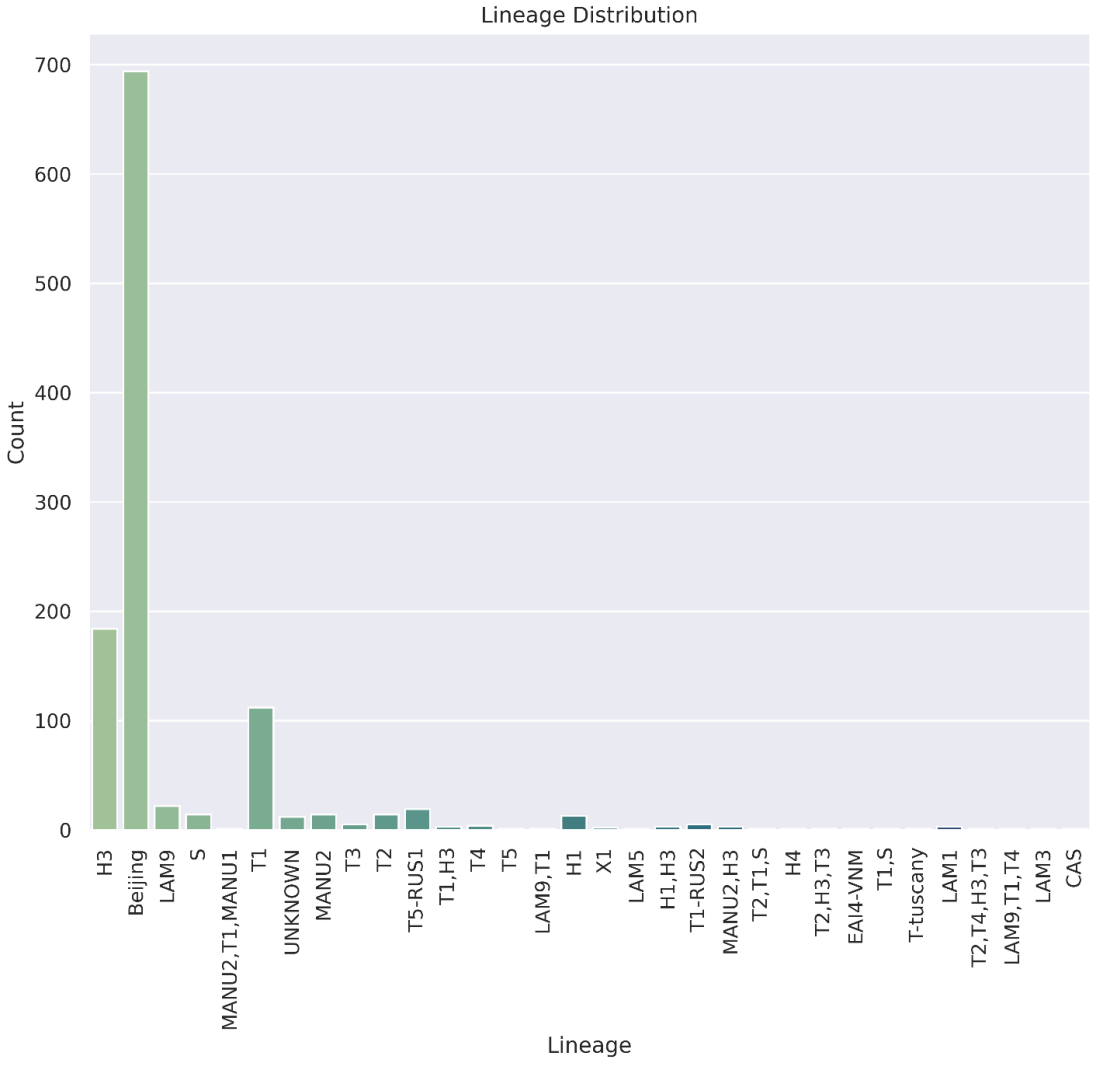

Type of Resistance
Count
700
600
500
400
300
200
100
0
MDR non XDR
Mono DR
Poly DR
Pre-XDR
XDR
Resistance

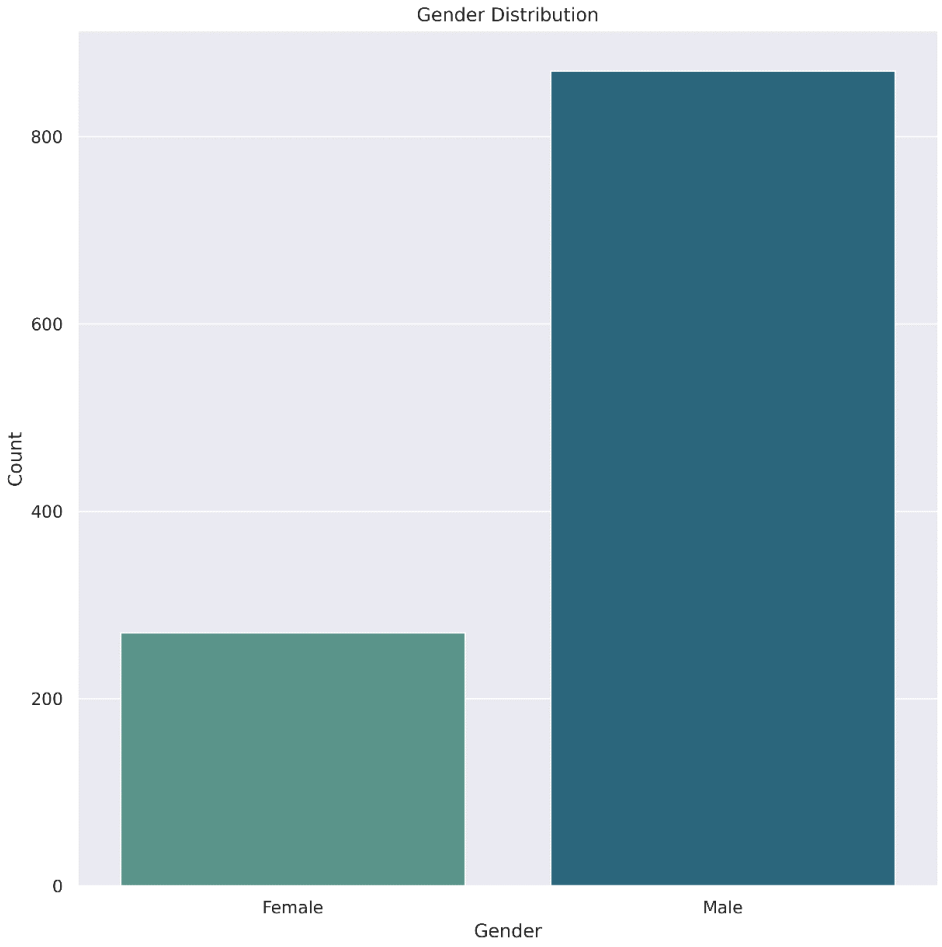

Type of Resistance vs Outcome Status
Count
600
500
400
300
200
100
0
MDR non XDR
Mono DR
Poly DR
Pre-XDR
Sensitive
XDR
Type of resistance
Cured
Died
Failure
Lost to follow up
Still on treatment
Completed
Unknown

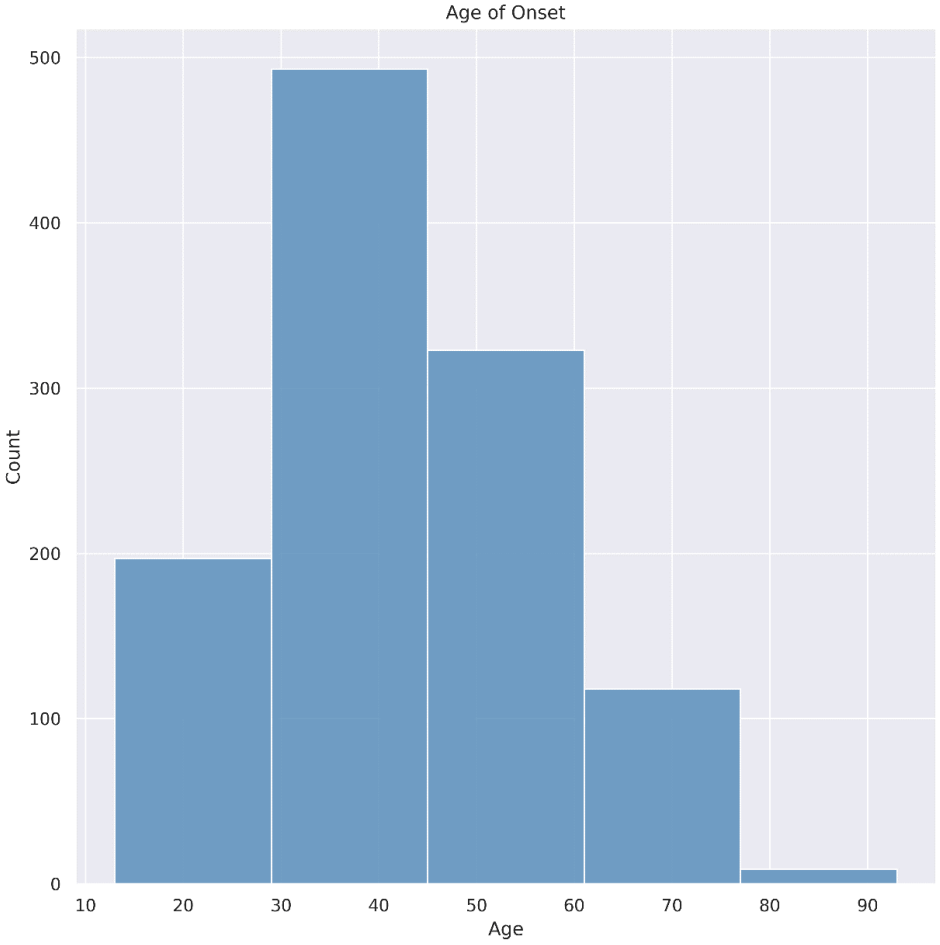

Type of Resistance vs Outcome Status
Cured
Died
Failure
Lost to follow up
Still on treatment
Completed
Unknown
Count
600
500
400
300
200
100
0
MDR non XDR
Mono DR
Poly DR
Pre-XDR
Sensitive
XDR
Type of resistance

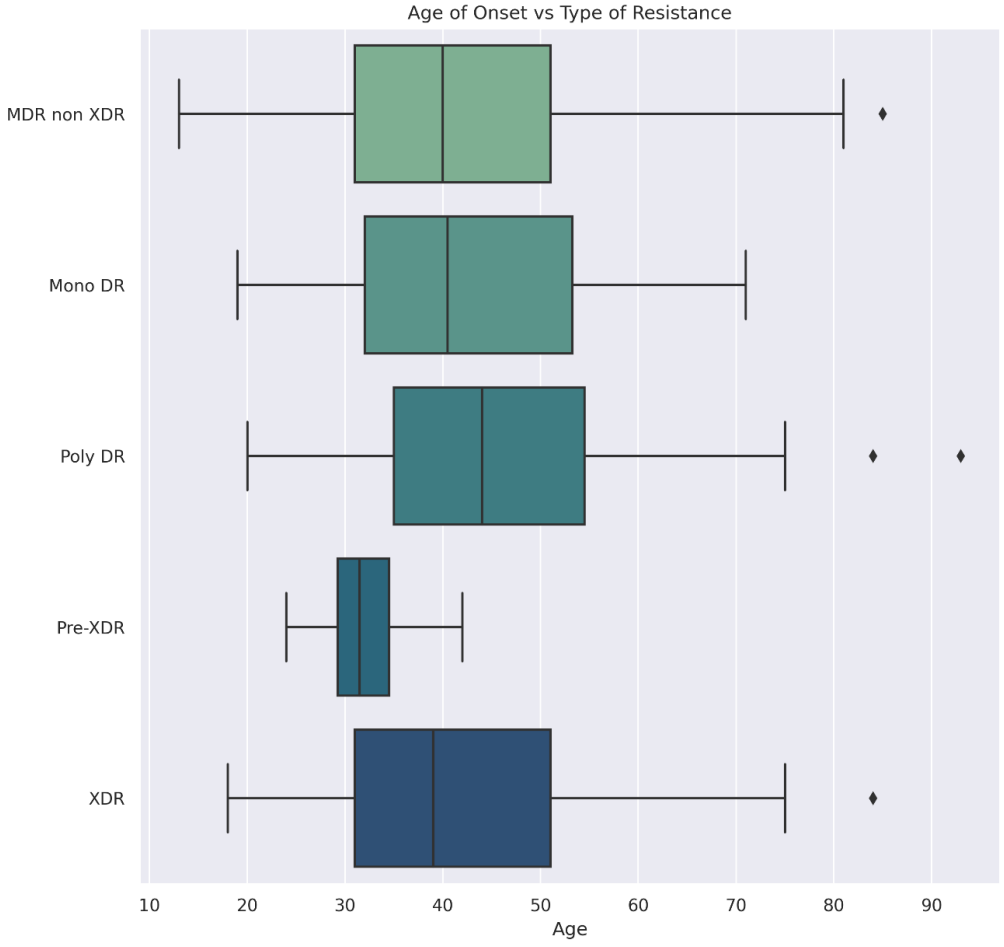

Lineage vs Outcome Status
Died
Failure
Lost to follow up
Still on treatment
Completed
Unknown
Count
0
20
40
60
80
100
120
140
160
H3
Beijing
LAM9
S
MANU2,T1,MANU1
T1
UNKNOWN
MANU2
T3
T2
T5-RUS1
T1,H3
T4
T5
LAM9,T1
H1
X1
LAM5
H1,H3
T1-RUS2
MANU2,H3
T2,T1,S
H4
T2,H3,T3
EAI4-VNM
T1,S
T-tuscany
LAM1
T2,T4,H3,T3
LAM9,T1,T4
LAM3
CAS
Lineage

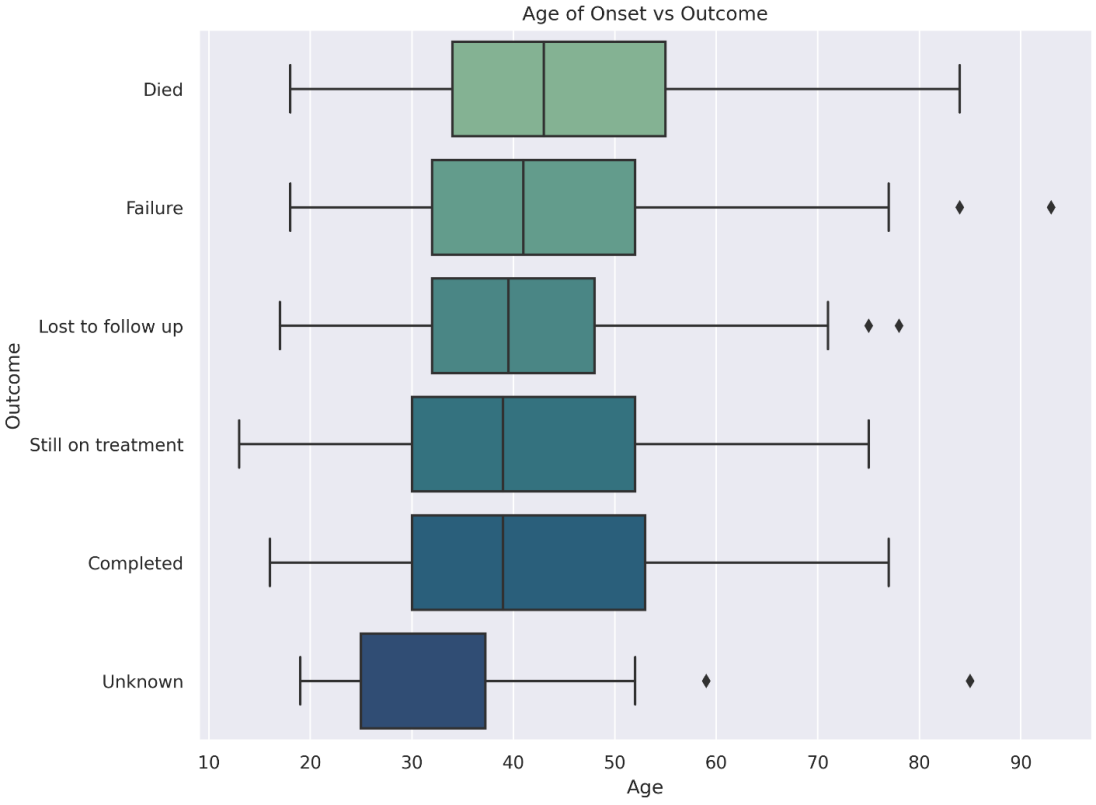

Outcome
Count
250
200
150
100
50
0
Died
Failure
Lost to follow up
Still on treatment
Completed
Unknown
Outcomes

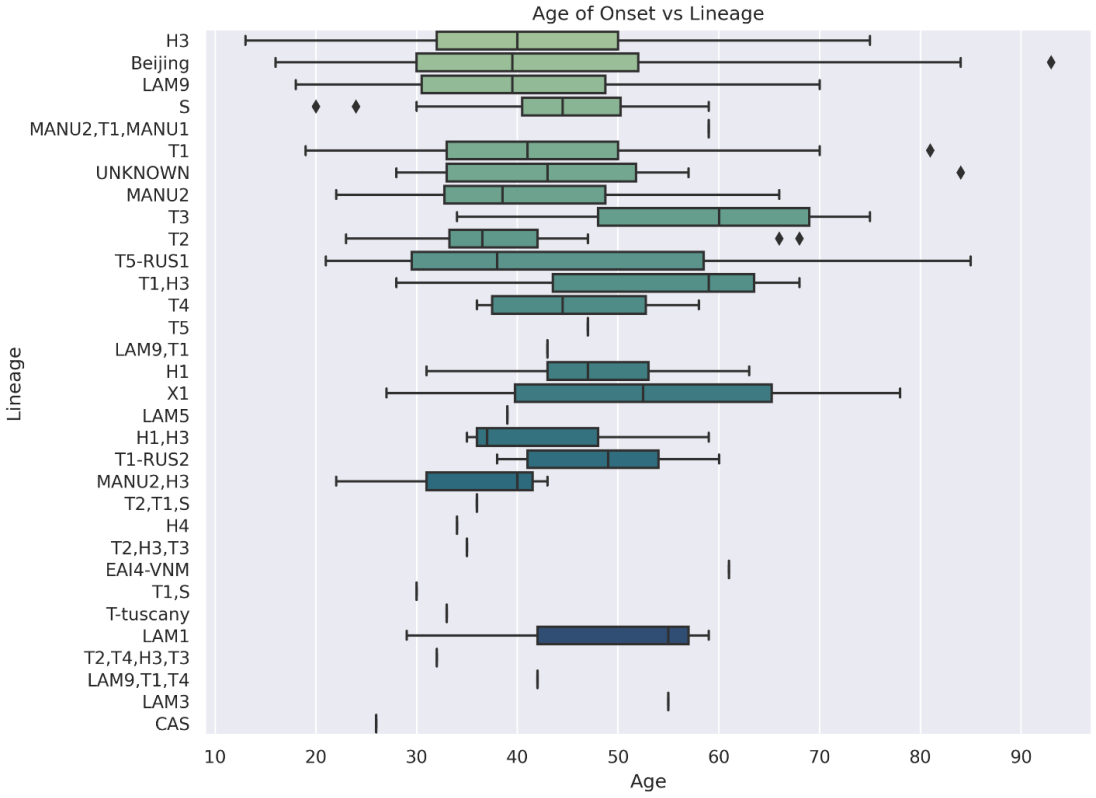

Lineage Distribution
Count
0
100
200
300
400
500
600
700
H3
Beijing
LAM9
S
MANU2,T1,MANU1
T1
UNKNOWN
MANU2
T3
T2
T5-RUS1
T1,H3
T4
T5
LAM9,T1
H1
X1
LAM5
H1,H3
T1-RUS2
MANU2,H3
T2,T1,S
H4
T2,H3,T3
EAI4-VNM
T1,S
T-tuscany
LAM1
T2,T4,H3,T3
LAM9,T1,T4
LAM3
CAS
Lineage


Gender Distribution
800
600
400
200
0
Count
Female
Male
Gender


Age of Onset
Count
0
100
200
300
400
500
10
20
30
40
50
60
70
80
90
Age


Age of Onset vs Type of Resistance
MDR non XDR
Mono DR
Poly DR
Pre-XDR
XDR
10
20
30
40
50
60
70
80
90
Age


Age of Onset vs Outcome
Outcome
Died
Failure
Lost to follow up
Still on treatment
Completed
Unknown
10
20
30
40
50
60
70
80
90
Age


Age of Onset vs Lineage
Lineage
H3
Beijing
LAM9
S
MANU2,T1,MANU1
T1
UNKNOWN
MANU2
T3
T2
T5-RUS1
T1,H3
T4
T5
LAM9,T1
H1
X1
LAM5
H1,H3
T1-RUS2
MANU2,H3
T2,T1,S
H4
T2,H3,T3
EAI4-VNM
T1,S
T-tuscany
LAM1
T2,T4,H3,T3
LAM9,T1,T4
LAM3
CAS
10
20
30
40
50
60
70
80
90
Age

Appendix E

.5 Association Rules above support 0.01

Antecedent		Consequent	Support	Conf.
RIF, INH	=>	CAP, OFL, STR	0.07614213	1
RIF, INH	=>	CAP, EMB, OFL, STR	0.07005076	1
RIF,INH	=>	OFL, AMK, STR	0.05989848	1
RIF,INH	=>	OFL, EMB, AMK, STR	0.05583756	1
RIF,INH	=>	KAN, CAP, OFL, STR	0.05279188	1
RIF,INH	=>	STR, CAP, AMK, OFL	0.05228426	1
RIF,INH	=>	KAN, CAP, OFL, STR, EMB	0.05076142	1
RIF,INH	=>	AMK, CAP, OFL, STR, EMB	0.04923858	1
RIF,INH	=>	KAN, OFL, AMK, STR	0.04162437	1
RIF,INH	=>	KAN, AMK, OFL, STR, EMB	0.04060914	1
RIF,INH	=>	KAN, AMK, CAP, OFL, STR	0.03959391	1
RIF,INH	=>	KAN, AMK, CAP, OFL, STR, EMB	0.03908629	1
RIF,INH	=>	CAP, EMB, PTH, STR	0.02588833	1
RIF,INH	=>	CS, EMB	0.02233503	1
RIF,INH	=>	EMB, PTH, AMK, STR	0.02233503	1
RIF,INH	=>	CAP, LVX, STR	0.02182741	1
RIF,INH	=>	KAN, PZA, EMB	0.0213198	1
RIF,INH	=>	KAN, PTH, STR	0.0213198	1
RIF,INH	=>	CAP, PTH, OFL, STR	0.02081218	1
RIF,INH	=>	LVX, AMK, STR	0.02030457	1
RIF,INH	=>	AMK, CAP, STR, PTH, EMB	0.02030457	1
RIF,INH	=>	CS, EMB, OFL	0.01928934	1
RIF,INH	=>	KAN, EMB, PTH, STR	0.01928934	1
RIF,INH	=>	CS, STR	0.01878173	1
RIF,INH	=>	LVX, EMB, AMK, STR	0.01878173	1
RIF,INH	=>	CAP, EMB, LVX, STR	0.01878173	1
RIF,INH	=>	CAP, OFL, STR, PTH, EMB	0.01878173	1
RIF,INH	=>	KAN, CS, EMB	0.01827411	1
RIF,INH	=>	LVX, EMB, OFL	0.01827411	1
RIF,INH	=>	KAN, CAP, PTH, STR	0.01827411	1
RIF,INH	=>	CS, EMB, STR	0.0177665	1

Antecedent		**Consequent**	**Support**	**Conf.**
RIF,INH	=>	LVX, EMB, OFL, STR	0.0177665	1
RIF,INH	=>	OFL, PTH, AMK, STR	0.01725888	1
RIF,INH	=>	KAN, CS, EMB, OFL	0.01675127	1
RIF,INH	=>	AMK, OFL, STR, PTH, EMB	0.01675127	1
RIF,INH	=>	KAN, CAP, STR, PTH, EMB	0.01675127	1
RIF,INH	=>	CS, OFL, STR	0.01624366	1
RIF,INH	=>	CAP, LVX, AMK, STR	0.01624366	1
RIF,INH	=>	KAN, PTH, AMK, STR	0.01573604	1
RIF,INH	=>	KAN, PTH, OFL, STR	0.01573604	1
RIF,INH	=>	KAN, EMB, PAS	0.01522843	1
RIF,INH	=>	CS, EMB, OFL, STR	0.01522843	1
RIF,INH	=>	KAN, AMK, CAP, STR, PTH	0.01522843	1
RIF,INH	=>	AMK, CAP, OFL, STR, PTH	0.01522843	1
RIF,INH	=>	KAN, AMK, STR, PTH, EMB	0.01522843	1
RIF,INH	=>	KAN, CS, STR	0.01472081	1
RIF,INH	=>	KAN, PZA, STR	0.01472081	1
RIF,INH	=>	AMK, CAP, STR, LVX, EMB	0.01472081	1
RIF,INH	=>	KAN, OFL, STR, PTH, EMB	0.01472081	1
RIF,INH	=>	KAN, AMK, CAP, STR, PTH, EMB	0.01472081	1
RIF,INH	=>	AMK, CAP, OFL, STR, PTH, EMB	0.01472081	1
RIF,INH	=>	OFL, LVX, AMK	0.0142132	1
RIF,INH	=>	PAS, CAP, EMB	0.0142132	1
RIF,INH	=>	CAP, LVX, OFL	0.0142132	1
RIF,INH	=>	CAP, LVX, OFL, STR	0.0142132	1
RIF,INH	=>	KAN, CS, EMB, STR	0.0142132	1
RIF,INH	=>	KAN, PZA, EMB, STR	0.0142132	1
RIF,INH	=>	KAN, CAP, OFL, STR, PTH	0.0142132	1
RIF, INH, STR	=>	CAP, LVX, OFL	0.0142132	1
RIF, INH, STR	=>	CAP, LVX, OFL	0.0142132	1
RIF, INH, STR	=>	CAP, LVX, OFL	0.0142132	1
RIF,INH	=>	CS, CAP, EMB	0.01370558	1
RIF,INH	=>	OFL, LVX, AMK, STR	0.01370558	1
RIF,INH	=>	KAN, CS, CAP, EMB	0.01370558	1
RIF,INH	=>	KAN, CS, OFL, STR	0.01370558	1
KAN, RIF, INH	=>	CS, CAP, EMB	0.01370558	1
KAN, RIF, INH	=>	CS, CAP, EMB	0.01370558	1
KAN, RIF, INH	=>	CS, CAP, EMB	0.01370558	1
RIF,INH	=>	CS, EMB, AMK	0.01319797	1
RIF,INH	=>	PZA, EMB, OFL	0.01319797	1
RIF,INH	=>	LVX, PTH, STR	0.01319797	1
RIF,INH	=>	LVX, OFL, EMB, AMK	0.01319797	1
RIF,INH	=>	CS, CAP, EMB, OFL	0.01319797	1
RIF,INH	=>	KAN, CAP, OFL, CS, EMB	0.01319797	1
RIF,INH	=>	KAN, OFL, STR, CS, EMB	0.01319797	1
RIF,INH	=>	KAN, CAP, OFL, STR, PTH, EMB	0.01319797	1
KAN, RIF, INH	=>	CS, CAP, EMB, OFL	0.01319797	1
KAN, RIF, INH	=>	CS, CAP, EMB, OFL	0.01319797	1
KAN, RIF, INH	=>	CAP, OFL, CS, EMB	0.01319797	1

Antecedent		**Consequent**	**Support**	**Conf.**
RIF,INH	=>	PZA, OFL, STR	0.01269036	1
RIF,INH	=>	KAN, CS, EMB, AMK	0.01269036	1
RIF,INH	=>	CS, OFL, EMB, AMK	0.01269036	1
RIF,INH	=>	PZA, EMB, OFL, STR	0.01269036	1
RIF,INH	=>	AMK, OFL, STR, LVX, EMB	0.01269036	1
EMB, RIF, INH	=>	PZA, OFL, STR	0.01269036	1
RIF, EMB, INH	=>	PZA, OFL, STR	0.01269036	1
RIF, INH, EMB	=>	PZA, OFL, STR	0.01269036	1
RIF,INH	=>	CS, CAP, EMB, AMK	0.01218274	1
RIF,INH	=>	LVX, CAP, EMB, OFL	0.01218274	1
RIF,INH	=>	PAS, CAP, EMB, OFL	0.01218274	1
RIF,INH	=>	KAN, AMK, CAP, CS, EMB	0.01218274	1
RIF,INH	=>	KAN, AMK, OFL, CS, EMB	0.01218274	1
RIF,INH	=>	CAP, OFL, STR, LVX, EMB	0.01218274	1
KAN, RIF, INH	=>	CS, CAP, EMB, AMK	0.01218274	1
RIF, INH, STR	=>	LVX, CAP, EMB, OFL	0.01218274	1
KAN, RIF, INH	=>	CS, CAP, EMB, AMK	0.01218274	1
KAN, RIF, INH	=>	AMK, CAP, CS, EMB	0.01218274	1
RIF, INH, STR	=>	LVX, CAP, EMB, OFL	0.01218274	1
RIF, INH, STR	=>	CAP, OFL, LVX, EMB	0.01218274	1
RIF,INH	=>	CS, AMK, STR	0.01167513	1
RIF,INH	=>	CS, CAP, STR	0.01167513	1
RIF,INH	=>	KAN, CAP, EMB, PAS	0.01167513	1
RIF,INH	=>	KAN, EMB, PAS, STR	0.01167513	1
RIF,INH	=>	AMK, CAP, OFL, CS, EMB	0.01167513	1
RIF,INH	=>	KAN, AMK, OFL, STR, PTH	0.01167513	1
RIF,INH	=>	KAN, AMK, CAP, OFL, CS, EMB	0.01167513	1
RIF,INH	=>	KAN, AMK, OFL, STR, PTH, EMB	0.01167513	1
EMB, RIF, INH	=>	KAN, AMK, OFL, STR, PTH	0.01167513	1
KAN, RIF, INH	=>	AMK, CAP, OFL, CS, EMB	0.01167513	1
KAN, RIF, INH	=>	AMK, CAP, OFL, CS, EMB	0.01167513	1
KAN, RIF, INH	=>	AMK, CAP, OFL, CS, EMB	0.01167513	1
RIF, EMB, INH	=>	KAN, AMK, OFL, STR, PTH	0.01167513	1
RIF, EMB, INH	=>	KAN, AMK, OFL, STR, PTH	0.01167513	1
RIF,INH	=>	CAP, LVX, AMK, OFL	0.01116751	1
RIF,INH	=>	CS, EMB, AMK, STR	0.01116751	1
RIF,INH	=>	CS, OFL, AMK, STR	0.01116751	1
RIF,INH	=>	KAN, CS, CAP, STR	0.01116751	1
RIF,INH	=>	CS, CAP, OFL, STR	0.01116751	1
RIF,INH	=>	PAS, CAP, EMB, STR	0.01116751	1
RIF,INH	=>	AMK, CAP, OFL, STR, LVX	0.01116751	1
RIF,INH	=>	KAN, AMK, CAP, OFL, STR, PTH	0.01116751	1
RIF,INH	=>	KAN, AMK, CAP, OFL, STR, PTH, EMB	0.01116751	1
RIF, EMB, INH	=>	KAN, AMK, CAP, OFL, STR, PTH	0.01116751	1
RIF, EMB, INH	=>	KAN, AMK, CAP, OFL, STR, PTH	0.01116751	1
EMB, RIF, INH	=>	KAN, AMK, CAP, OFL, STR, PTH	0.01116751	1
RIF, INH, STR	=>	CAP, LVX, AMK, OFL	0.01116751	1

Antecedent		**Consequent**	**Support**	**Conf.**
RIF, INH, STR	=>	CAP, LVX, AMK, OFL	0.01116751	1
RIF, INH, STR	=>	AMK, CAP, OFL, LVX	0.01116751	1
RIF,INH	=>	CS, EMB, PTH	0.0106599	1
RIF,INH	=>	CS, CAP, AMK, STR	0.0106599	1
RIF,INH	=>	KAN, CS, AMK, STR	0.0106599	1
RIF,INH	=>	CS, CAP, EMB, STR	0.0106599	1
RIF,INH	=>	LVX, EMB, PTH, STR	0.0106599	1
RIF,INH	=>	KAN, AMK, STR, CS, EMB	0.0106599	1
RIF,INH	=>	AMK, OFL, STR, CS, EMB	0.0106599	1
RIF,INH	=>	KAN, CAP, STR, CS, EMB	0.0106599	1
RIF,INH	=>	KAN, CAP, OFL, STR, CS	0.0106599	1
RIF, EMB, INH	=>	KAN, CS, AMK, STR	0.0106599	1
RIF, EMB, INH	=>	KAN, AMK, STR, CS	0.0106599	1
KAN, RIF, INH	=>	CS, CAP, EMB, STR	0.0106599	1
KAN, RIF, INH	=>	CAP, STR, CS, EMB	0.0106599	1
EMB, RIF, INH	=>	KAN, CS, AMK, STR	0.0106599	1
KAN, RIF, INH	=>	CS, CAP, EMB, STR	0.0106599	1
RIF,INH	=>	PZA, EMB, AMK	0.01015228	1
RIF,INH	=>	CAP, PZA, EMB	0.01015228	1
RIF,INH	=>	PAS, CAP, OFL, STR	0.01015228	1
RIF,INH	=>	KAN, EMB, PAS, OFL	0.01015228	1
RIF,INH	=>	AMK, CAP, STR, CS, EMB	0.01015228	1
RIF,INH	=>	KAN, AMK, CAP, STR, CS	0.01015228	1
RIF,INH	=>	AMK, CAP, OFL, STR, CS	0.01015228	1
RIF,INH	=>	AMK, CAP, OFL, LVX, EMB	0.01015228	1
RIF,INH	=>	KAN, AMK, OFL, STR, CS	0.01015228	1
RIF,INH	=>	CAP, OFL, STR, CS, EMB	0.01015228	1
RIF,INH	=>	KAN, AMK, CAP, STR, CS, EMB	0.01015228	1
RIF,INH	=>	AMK, CAP, OFL, STR, LVX, EMB	0.01015228	1
RIF,INH	=>	KAN, AMK, OFL, STR, CS, EMB	0.01015228	1
RIF,INH	=>	KAN, CAP, OFL, STR, CS, EMB	0.01015228	1
RIF, EMB, INH	=>	KAN, AMK, CAP, STR, CS	0.01015228	1
RIF, EMB, INH	=>	KAN, AMK, OFL, STR, CS	0.01015228	1
RIF, EMB, INH	=>	KAN, AMK, CAP, STR, CS	0.01015228	1
RIF, EMB, INH	=>	KAN, AMK, OFL, STR, CS	0.01015228	1
RIF, INH, STR	=>	AMK, CAP, OFL, LVX, EMB	0.01015228	1
RIF, INH, STR	=>	AMK, CAP, OFL, LVX, EMB	0.01015228	1
KAN, RIF, INH	=>	AMK, CAP, STR, CS, EMB	0.01015228	1
KAN, RIF, INH	=>	CAP, OFL, STR, CS, EMB	0.01015228	1
KAN, RIF, INH	=>	AMK, CAP, STR, CS, EMB	0.01015228	1
KAN, RIF, INH	=>	, CAP, OFL, STR, CS, EMB	0.01015228	1
KAN, RIF, INH	=>	AMK, CAP, STR, CS, EMB	0.01015228	1
KAN, RIF, INH	=>	CAP, OFL, STR, CS, EMB	0.01015228	1
RIF, INH, STR	=>	AMK, CAP, OFL, LVX, EMB	0.01015228	1
EMB, RIF, INH	=>	KAN, AMK, CAP, STR, CS	0.01015228	1
EMB, RIF, INH	=>	KAN, AMK, OFL, STR, CS	0.01015228	1